REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2005 Name of Author ANDREW, A.

Standardisation of near infrared spectrophotometers

Anne Andrew

University College London

Submitted for the degree of Doctor of Philosophy

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

A near infrared (NIR) spectrometer produces, from a single sample, a spectrum formed from several hundred absorbance readings at a range of wavelengths in the NIR region. Using regression approaches and a large number of samples for which reference values and spectra are known, the instrument can be calibrated to predict reference values from spectra. A problem with NIR spectrometers is that no two instruments produce exactly the same output, as a result of which a calibration developed on one instrument cannot be transferred to a second instrument unless the second instrument has been standardised first.

Our aim in this thesis is to explore and assess improved methods of standardising NIR spectrometers. The main line of attack is to use standard models but incorporate prior information through Bayesian techniques. The main commercially used standardisation techniques adjust the spectra wavelength by wavelength without any use being made of the fact that the spectra and therefore the appropriate adjustment varies smoothly. By the use of suitable priors within a Bayesian analysis we produce a better solution. The analysis is very time-consuming, involving inverting large matrices and MCMC or some other process for determining parameters. A second attempt using the same assumptions uses dynamic linear modelling, treating the spectra as time series. While theoretically slightly inferior, this method is very much quicker and produces comparable results. A third solution, while using the same basic model, makes an estimate of the wavelength shift in the wavelet domain. Our final, non-Bayes, method is intended to standardise a number of similar instruments

simultaneously. This is achieved by projecting spectra onto a subspace orthogonal to the space spanned by between-instrument variation and calibrating on the subspace to produce a robust calibration.

Acknowledgements

I would like to thank my supervisor, Professor Tom Fearn, for persuading me to embark on a research degree and for his continuing encouragement, guidance and endless patience. His knowledge and inspiration and the fact that he was always prepared to spend time in discussion were invaluable to me. I would also like to thank Dr. Tim Downie for his advice and suggestions. I am most grateful to my examiners, Professor David Hand and Dr. Pierre Dardenne for their interest and helpful criticism.

The friendship, support and encouragement of my fellow research students has been of enormous value to me.

My deepest gratitude goes to my husband, Stephen, whose understanding and enthusiasm for my irrational desire to study statistics have never faltered.

Contents

List of Figures

List of Tables

Notation

I: identity matrix.

I_n: $n \times n$ identity matrix.

1_n: $n \times 1$ vector of ones.

X^T: the transpose of the vector or matrix X.

$|X|$: the determinant of matrix X.

X^{-1}: the inverse of matrix X.

X^-: a generalised inverse of X.

$X(i,j)$ or x_{ij}: the element in the ith row and the jth column of X.

$X(i)$ or x_i: the ith element of the vector X.

$X(:,j)$: the jth column of X, as a vector.

$\bar{X}_{i.}$: the mean of the elements in the ith row of X.

$\bar{X}_{..}$: the mean of all the elements of X.

$[a\ b\ \ldots]$: the matrix whose columns are $a, b, \ldots$

$X \otimes Y$: the Kroneker or direct product of matrices X and Y:

$$X \otimes Y = \begin{pmatrix} X(1,1)Y & X(1,2)Y & \ldots & \ldots \\ X(2,1)Y & X(2,2)Y & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

<a,b, ... > : the vector space spanned by $a, b, \ldots$.

P(.): the probability of the event contained in parentheses.

|: 'conditional on'

E(T): the expectation of the random variable T.

var(T): the variance of the random variable T.

$N(\mu, \Sigma)$: a normal distribution with expectation μ and variance-covariance matrix Σ

$\mathcal{N}(\Sigma, \Gamma)$: a matrix-normal distribution. $\gamma_{ii}\Sigma$ represents the variance-covariance matrix of the ith column and $\sigma_{jj}\Gamma$ of the jth row of the matrix.

$ga(\alpha, \beta)$: a gamma distribution with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$

Chapter 1

Introduction

1.1 Near infrared spectrometers

It has been known since the beginning of the 19th century that energy exists beyond the red end of the visible spectrum in the area named the infrared region by Sir William Herschel who first noticed this phenomonen. A century later W. W. Coblentz built a spectrophotometer or spectrometer and used it to measure the spectra of several hundred compounds. He realised that substances could be identified from their infrared spectra. Interest centred on the mid-infrared region of the spectrum where sharp peaks and the well-defined shapes of the spectra make identification easy. Only in the last 25 years with improved instruments and more sophisticated analytical techniques has the use of the near infrared (NIR) region of the spectrum, the part just beyond the visible range, become significant. The linear relationship between absorbance, the output of the spectrometer, and the concentration of a substance is exploited to provide an analysis of a substance; a thousand or more absorbance readings being combined to predict the concentrations of constituents in a substance. Near-infrared spectrometers can be programmed to produce results very rapidly and do not need highly skilled technicians to operate them. Consequently they have become increasingly important in the analysis of a

variety of substances in, for example, agriculture and the pharmaceutical and petrochemical industries.

Historically the most important use of NIR spectroscopy was in the analysis of wheat. NIR spectrometers are used to provide an almost instantaneous assessment of the quality of wheat, allowing rapid decisions to be made about the suitability of samples of wheat for different uses. NIR spectrometers are also used to analyse other agricultural products and more generally in the food industry. The speed with which the analysis is performed has made possible their use for on line control in food manufacturing. The advantages of using NIR spectroscopy clearly apply to its use in the pharmaceutical industry. However the industry has been slow to implement spectroscopic methods. This has been partly due to regulatory controls which are written in terms of chromatographic methods. Research has shown that NIR spectroscopic methods for quality control and quantitative analysis can meet regulatory standards.

Multiple regression or some related method is used to calibrate the spectrometer to give an analysis of the constituents in a substance from the absorbances in the near infrared region. Calibration may involve several hundred samples with known reference values, collected over a considerable period of time. The samples are often perishable and not easily transportable. Consequently the process is both costly and time-consuming. It is therefore important to be able to transfer the calibration to other similar instruments without having to repeat the whole exercise. A problem arises here because the spectrum of any given sample will typically vary a little when measured on different instruments, or on the same instrument after it has been in use for some time or has undergone some change due to servicing or repair. The differences usually occur either because of imprecision in locating the position of the wavelength or because of a shift in the absorbance measurement, or both. As a result of this, a calibration developed on one instrument usually performs less well on another unless some adjustment is made. It is this problem of calibration

transfer that we address in this thesis.

One approach to calibration transfer is to develop calibrations that are robust to different instruments. A possible way of doing this is to adjust all spectra before calibrating. NIR spectra are often subjected to pretreatments before being used for calibration and prediction. These are generally intended to remove additive and multiplicative differences between spectra caused by differing particle sizes but will also remove between-instrument differences if these differences are of the same kind. More sophisticated methods for producing robust calibrations reduce the dimension of the calibration space by removing unwanted variation before calibration.

Many of the more successful methods of instrument standardisation use spectra from a few carefully chosen samples that have been measured on both the instrument on which the calibration was developed and on those instruments to which it is to be transferred. Clearly the number of samples used for standardisation should be as small as possible, and should be substantially smaller than the calibration set or re-calibration would be preferable. The most accurate methods of calibration transfer and those that are most are widely used commercially are methods that adjust the spectra of samples on a new instrument so that they match those of an already calibrated instrument. The existing calibration can then be transferred to the new instrument.

1.2 Bayesian methods

Bayesian methods provide not only a firm basis for statistical inference, offering information about parameters of interest in the form of probability distributions on the parameters, but also a method for combining prior knowledge with observed data to enable inferences to be made. Using Bayesian hierarchical models a structure can be imposed upon the data and parameters.

The traditional approach to estimating unknown parameters in a situation where data are available is to find the value of the parameter, θ, that maximises

the likelihood, $P(X|\theta)$, of θ, given the data, X. If prior information about the parameter in the form of a probability distribution, $P(\theta)$, exists then the Bayesian approach is to combine this with the likelihood to give a probability distribution for θ conditional on the data. The distribution $P(\theta|X)$ summarises information about the value of θ.

One situation where Bayesian methods offer advantages is in multiple regression. Using maximum likelihood estimation will fail to produce a result in the case where there are fewer samples than variables because the covariance matrix, which must be inverted, will be singular. Even if there are sufficient data the result may be unstable. Various methods have been developed to avoid the problem. Principle components regression (PCR) and partial least squares regression (PLS), special cases of continuum regression (Stone and Brooks (1990)), aim to reduce the the number of variables while retaining the useful information. In ridge regression the sample covariance matrix is stabilised by the addition of a diagonal matrix. A justification for this has been given by Lindley and Smith (1971) using a Bayesian approach. This provides a theoretical justification and also a framework within which generalisations can be made.

The Bayesian approach avoids the restriction caused by insufficient data, by using prior information. Prior information about regression coefficients is defined in the form of probability distributions for the coefficients. Structure may be imposed on the model by the use of appropriate prior distributions, for example, the coefficients might be modelled as independent and identically distributed or a multivariate distribution might be used, with a correlation structure defined by parameters of the distribution. The parameters may be specified or they may themselves be assumed to follow specified distributions, thus creating a hierarchy.

Bayesian hierarchical models of the kind just described may be extremely complex. It is only with the advent of increased computer power that it has

been possible to exploit their full potential. The most important technique available for tackling these complex Bayesian models is Markov chain Monte Carlo (MCMC), an iterative process, which involves sampling from a sequence of distributions constructed so that they eventually converge to the joint distribution of all the parameters.

1.3 Wavelets

Wavelets are increasingly being used to model or represent functions. A set of wavelets is formed by translation and dilation of a single compactly supported function, creating an orthogonal basis for a wide variety of functions of different degrees of smoothness. These properties allow the exact representation of a discretely sampled function in the wavelet domain. It is also possible to define a function at different levels of resolution providing a parsimonious representation in terms of wavelet coefficients. An important application of wavelets is denoising signals. Suppose a signal is sampled at equally spaced intervals but is observed with added white noise. In the wavelet domain, the signal will be represented as a linear combination of wavelets and white noise and in this representation, small coefficients represent noise rather than signal and can be ignored. This process is known as hard thresholding. An alternative to this, soft thresholding, involves setting small coefficients to zero but also shrinking larger coefficients either by a fixed amount or by an amount that varies with the level of resolution. Thresholding leads to a function from which noise has been eliminated and which is also efficiently represented. Thresholding can be performed within a Bayesian framework by placing appropriate priors on the wavelet coefficients. Abramovich et al. (1998) propose a form of prior for the wavelet coefficients that has the effect of thresholding the coefficients. They establish a relationship between the prior hyperparameters and the particular Besov space in which the resulting function will lie.

1.4 Application of Bayesian methods to the problem of standardisation of NIR spectrometers

Two of the most successful methods of standardisation used commercially are the Shenk-Westerhaus patented method (SW) (Shenk and Westerhaus (1989), Bouveresse et al. (1994)) and Piecewise Direct Standardisation (PDS) (Wang et al. (1991)). In each of these the spectral adjustment is found by regressing absorbances from one instrument at a given wavelength on those for the same samples measured on another instrument, either at the same wavelength or on a small window centred on that wavelength. With a small standardisation set this leads to the usual problems of over-fitting and instability. The standard Bayesian approach to regression is to place a prior on the regression coefficients and use Bayes' theorem to determine posterior distributions for the coefficients. In this particular situation we have information that can be used to improve the solution. We can use the fact that the differences between the two instrumental responses are likely to be small to select a prior mean for the regression coefficients and the fact that the regression coefficients will vary smoothly from wavelength to wavelength allows us to impose a structure on them.

In chapter 5 we use the same models for the data as are used commercially and develop Bayesian posterior distributions. We apply a variety of techniques to determine the associated parameters. The most successful of these involves constructing a Bayesian hierarchical model and using MCMC to determine the posterior distrubution of the parameters of the model.

The Bayesian approach, because it involves inverting large matrices and an iterative procedure to evaluate parameters requires a large amount of computer time. In chapter 7 we investigate the use of dynamic linear modelling, an approach described in the literature of time series. Here the absorbances at consecutive wavelengths are treated as a time series. As for SW and PDS a

linear relationship between slave and master instruments is assumed. Each regression coefficient is linked to the succeeding one by a system equation, the effect of which is to allow only small variation between adjacent regression coefficients and to shrink coefficients towards the null assumption that slave and master instrumental responses are the same. We again use the models that are currently used, and make similar prior assumptions to those made in chapter 5. This approach is far quicker and produces comparable results.

Currently used models fail to produce accurate solutions. One reason for this is that the correlations between absorbances at adjacent wavelengths lead to over-fitting. Although the methods described in chapters 5 and 7 reduce these effects, that they are still present is evident from the regression coefficients. In an attempt to overcome this problem, we used a different model and a new approach.

In chapter 9 we model the wavelength shift function in the wavelet domain. A prior was placed on the coefficients shrinking them towards zero and setting them equal to zero with probability varying with the level of resolution, leading to a parsimonious representation. MCMC failed to converge in this situation. We used instead iterated conditional mode and simulated annealing, methods which guaranteed convergence, though not necessarily to an optimum solution.

1.5 Transfer by orthogonal projection

Our final method, described in chapter 10 is different from those mentioned so far in that it aims to standardise several instruments simultaneously. It exploits a simple idea and although it cannot be used in all situations, has the advantage of being quick and easy to apply, and works very well when applicable. In this method the spectral data are projected onto a subspace orthogonal to the space in which the variation between instruments occurs. A calibration is developed on this subspace which is found to be robust not only to the instruments used in forming the subspace but also to other instruments

of the same type.

The main limitation of the method is that it will not work in situations where there is a wavelength shift between instruments. Its main advantage is that many instruments can be adjusted simultaneously, so that it is particularly appropriate when many similar instruments are in use. From a practical point of view this is probably our most successful method.

Chapter 2

NIR Spectrometers

2.1 Introduction

Here we give a very brief summary of the main facts concerning NIR absorbance and the methods by which this phenomenon is harnessed in the analysis of organic materials. A more detailed treatment of the chemistry can be found in a good analytical chemistry text book, for example Fifield and Kealey (1995). Textbooks on NIR analysis offer more detail on the theory and working of NIR spectrometers. Osborne et al. (1993) or Burns and Ciurczak (1992) give comprehensive treatments.

2.2 NIR absorbance

A substance subjected to electromagnetic radiation will absorb radiation in varying amounts at different wavelengths. Chemical bonds exist between atoms, the most important ones for NIR being between hydrogen and carbon, nitrogen or oxygen in organic materials. The bonds vibrate at defined frequencies, depending mainly on the particular atoms involved, but also on the molecules of which they form a part. A bond can only absorb radiation which resonates at the same or a related frequency, and in doing so it moves

to a higher energy state or level. At room temperature most molecules are in the ground state of vibration. The transition from this state to the first level is known as the fundamental transition. The wavelength of radiation for this transition is usually in the middle infrared region - 2500-50000 nm. Transitions to higher levels are possible, though with successively lower probabilities. These transitions - called overtones - usually have wavelengths in the NIR region 1100-2500 nm. The information relating absorption to wavelength is usually presented as a graph known as a spectrum. See for example figure 2.2.

Fundamental transitions appear as well-defined peaks in infrared spectra which has made possible the identification and structural analysis of organic materials from their infrared spectra. Spectra in the near infrared (NIR) region tend, especially in the case of complex organic substances, to be more complicated and less well-defined, consisting of overtones and combination bands (vibrations due to bonds which are contiguous within atoms, or different types of vibration of the same bonds) which are broader and lower and often overlap and hence lack the prominent features of the infrared spectra, making interpretation less straight-forward.

Advances in technology which have produced more sensitive detectors have enhanced the use of NIR spectrometers. The other important factor in the increasing use of NIR spectrometers has been the application of multiple regression to the spectral output. The first published work applying this technique appeared in a paper by Ben-Gera and Norris (1968) in which multiple regression was used in the analysis of agricultural products. For a more detailed historical perspective see Burns and Ciurczak (1992).

2.2.1 Relation between absorbance and concentration

The use of multiple regression is based on the fact that the absorbance of a substance, the log of the reciprocal of the proportion of light that is not ab-

sorbed, is proportional to the concentrations of constituents of the substance. To show this the Beer-Lambert law is invoked. This states that the proportion of radiation absorbed by a substance is equal in equal thicknesses of an absorbing medium (Lambert (1760)) and is proportional to the concentration of that substance within the medium. From Beer-Lambert's law we have,

$$-dI/I \propto dn$$

where I is the intensity of incident light and n is the number of absorbing molecules. $n \propto l \times c$, where l represents the pathlength of incident light and c concentration of absorbing molecules
Integrating we get

$$\log(I_0/I) \propto l \times c \tag{2.1}$$

where I_0 is the intensity of the incident light and I the intensity of the light after it has passed through or been reflected by the medium. Light which is not absorbed is either transmitted or reflected. The quantity $R = I/I_0$ is known as the transmittance for transmitted light or reflectance for reflected light.

In a NIR spectrometer reflectance relative to a non-absorbing standard - for example a ceramic disc - rather than absolute reflectance is measured. To allow for this an intercept is included in the relationship (2.1) giving

$$\log(1/R) \propto l_0 + lc \tag{2.2}$$

$\log(1/R)$ is known as the absorbance. Because of the above relationship, the output of a NIR spectrometer is usually the absorbance. Our spectral data is always given as absorbances.

Because of the complex nature of the NIR spectrum, the absorbance at a given wavelength will be the combination of absorbances due to different constituents so that absorbance, $\log(1/R_i)$, at wavelength i is given by

$$\log(1/R_i) = k_{i0} + \sum_j k_{ij}c_j \qquad i = 1 \ldots w \tag{2.3}$$

Here k_{i0} is a constant and k_{ij} is proportional to the absorptivity at wavelength i of constituent C_j which has concentration c_j.

2.2.2 Kubelka-Munk derivation

Several assumptions have been made in deriving this relationship, in particular, the effect of scatter of light within the medium has been ignored. Light scattering occurs when the refractive index within a substance changes. Scatter depends on the surfaces within a substance through which light passes so is affected by particle size and by the surrounding medium. In their paper Kubelka and Munk (1931) derived a relationship which takes into account scatter within the medium:-

$$\frac{(1-R_\infty)^2}{2R_\infty} = \frac{K}{S}$$

where K, the absorption coefficient, is proportional to the concentration of absorbing molecules, S is the scatter coefficient and R_∞ is the reflectance in a sample of "infinite depth", i.e. a sample which is sufficiently deep so that no further absorption can take place. Experimental results have shown this relationship appears often to be less valid in practice than the previous, simpler, one. (Burns and Ciurczak (1992)). What is clear from the Kubelka-Munk relationship is that increased scatter will increase the proportion of light absorbed, so that, for example, larger particle size will result in increased absorbance.

2.3 NIR Spectrometers

The amount of light absorbed by a sample cannot be measured directly, but the light that is not absorbed can be.

A spectrometer consists essentially of a light source, a monochromator or interferometer, which separates the light from the source into light in a narrow band of wavelengths (monochromatic light), and a detector. The earliest

monochromators consisted of lenses which focused the incident light on a prism which split the light into separate wavelengths. Prism monchromators have now given way to diffraction grating or reflection monochromators. The theory of these is based on the wavelike nature of light. In a diffraction grating monochromator light is reflected off a surface engraved with parallel wedge-shaped grooves. At certain angles the lengths of paths of light will differ by integral multiples of a particular wavelength so that waves of this wavelength will interfere constructively, producing monochromatic light with this wavelength.

In an interferometer light is split into two beams which are recombined after one has been delayed by an amount δ which varies with time. This results in light at different wavelengths being produced. In an interferometer the transmitted or absorbed light is converted to a digital signal and analysed using fast Fourier transforms. Wavelength inaccuracy which is a problem in monochromaters is less evident in Fourier transform instruments. Grating and interferometer instruments are the most commonly used instruments, but others exist, depending on different technologies and designed for different applications.

Spectrometers are designed so that all unabsorbed light is either reflected (in a reflectance spectrometer) or transmitted (transmittance spectrometer). Light passes from a source through a monochrometor or an interferometer to a sample where some of the light is absorbed and the remainder either passes through the sample to a detector or is reflected to reach detectors placed obliquely.

2.4 Calibration of NIR spectrometers

As demonstrated in section 2.2.1 there is a relationship between absorbance and concentration which is, at least approximately, linear. The process of determining the equation linking absorbances to concentration is known as

calibration. The relationship 2.1 lacks a proportionality constant and even if it were possible to estimate this, the path followed by the light through the absorbing medium is extremely complicated so that accurate estimation of the pathlength, l, is unlikely to be possible. For this reason, rather than formulate and solve equations 2.3 directly, a linear relationship between concentration and absorbance is assumed and multiple regression is used to determine the regression coefficients. Because of the importance of NIR spectrometers in the analysis of organic substances, the use of multiple linear regression for their calibration has attracted much attention. Absorbances for a sample at a large number of wavelengths in the NIR range can be measured quickly and cheaply while concentrations of constituents if measured directly require costly and time-consuming laboratory techniques. By exploiting the relationship between absorbances at wavelengths in the NIR range and concentration almost instant predictions of concentration are possible. Since it was first used in the 1960s multivariate calibration has been refined and improved and new techniques introduced. Naes et al. (2002) is an excellent introduction to the main methods and developments.

2.4.1 Multiple linear regression

Calibration may involve several hundred samples collected over a considerable period of time. Spectra for each of the samples together with concentrations of the constituents of interest are required. We write $Y = (Y_1, Y_2, \ldots, Y_n)^T$ for the vector of concentrations of constituents of n samples and X for the $n \times w$ matrix of absorbances so that the jth column $X(:, j) = (X_{1j}, X_{2j}, \ldots, X_{nj})^T$ gives the absorbances of the n samples at the jth wavelength. X and Y constitute the training data from which the parameters of the regression model can be derived. Once these are determined the model can be used to predict concentrations, y, of future samples with known absorbances $x_1, x_2, \ldots, x_w$. We assume in what follows that X and Y have been mean-centred. There are

two possible regression models:-

the inverse model,

$$Y \sim N(X\beta^{(1)}, \Sigma_1)$$

and the classical model,

$$X^{(j)} \sim N(Y\beta_j^{(2)}, \Sigma_2)$$

$\beta^{(1)}$ and $\beta_j^{(2)}$ are $w \times 1$ vectors of regression coefficients, Σ_1 and Σ_2, $n \times n$ variance-covariance matrices.

Since the measured absorbances X are dependent on the concentrations of the samples the second of these two models appears to be the more appropriate. However by comparing the mean squared errors (MSE) for the predictors of y, in the asymptotic case, where the regression coefficients are assumed to be precisely estimated using the two methods, Berkson (1969) showed that in the univariate case the inverse method will give smaller MSE, as long as the predicted value, is no more than $\sqrt{2}$ standard deviations from the mean of the training data. Brown (1993) extended this result to the multivariate Y case. Here each of the components of the predicted value, after canonical transformation, is required to be within $\sqrt{2}$ standard deviations of the mean, so the result is less strong. In most NIR applications the training data are selected so that the range of responses is wider than will be encountered in practice. Consequently the inverse method is used in practice and it is the method used in this thesis.

Focussing on the inverse model, when $\Sigma_1 = I$, the maximum likelihood estimate of β is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

when the inverse, $(X^T X)^{-1}$, exists. $\hat{\beta}$ can be shown to be the best unbiased estimator of β that is linear in Y and it minimises the sum of squares of residuals $(Y - X\hat{\beta})^T (Y - X\hat{\beta})$. For this reason the method is usually known as least squares regression (LSR). These criteria are not necessarily the most appropriate. It has been shown (Hoerl and Kennard (1970)) that unbiasedness is

not essential and that minimising the sum of squares of residuals may produce estimates of β that vary considerably from the true value. Another problem with the maximum likelihood estimate is that if the matrix X^TX is singular it will not be invertible. This is always so if $n < w$ since X^TX will then not be full rank. In NIR applications X^TX is not normally full rank even when n is larger than w because of correlations between absorbances and consequently some form of variable reduction or regularisation is necessary. Even if X^TX is invertible the inverse may be unstable, in the sense that small perturbations in X may result in large fluctuations in $(X^TX)^{-1}$; this is the result of correlated columns in X.

In the following sections we describe methods that aim to overcome these problems.

2.4.2 Principal components regression

Principal components regression (PCR) is a variable reduction technique that aims to construct and use as regressors variables that explain most of the variance of X. In order to predict varying concentrations for different substances we need predictors that vary between substances. This is the motivation for principal components regression. Predictors are selected that maximise the between-sample variance. A $w \times 1$ vector, c, defines a linear combination Xc of the absorbances for the mean-centred data, X. c is selected to maximise the variance, c^TX^TXc. It can be shown (for example Stone and Brooks (1990)) that the sequence of orthogonal vectors, $c_1, c_2, \ldots$, that maximise the variance, subject to the constraint that each is orthogonal to those preceeding it and each has modulus 1, is the sequence of eigenvectors of X^TX, ordered according to the magnitude of the corresponding eigenvalues. The eigenvalues are the variances of the corresponding scores over the data. The c_i are known as the loadings of the process. The regressors, $t_i = Xc_i$, are known as scores and can be shown to be the unnormalised eigenvectors of XX^T. There are in theory

$min(n-1, w-1)$ distinct scores, but only a few with largest variance are used in the regression. Some kind of stopping rule is used to determine the number of regressors. Cross-validation is often used. (See section 2.4.4).

The decomposition of X for PCR can be summarised by the identity

$$X = t_1 c_1^T + t_2 c_2^T + \ldots + t_r c_r^T + E. \tag{2.4}$$

Here E is a residual matrix.

PCR often works well in practice but there is no guarantee that the scores with large variance are necessarily the best predictors of concentration. In NIR spectroscopy it may well be that the large variation is due to scatter and hence reflects particle size rather than variation in concentration. In one of our examples we find that large variance is due to factors that are entirely uncorrelated with concentration, leading to scores that are poor predictors.

2.4.3 Partial least squares

Partial least squares (PLS), like PCR, constructs as regressors linear combinations of the spectra. For PLS the first loading c_1 is chosen so that it maximises $c^T X^T Y Y^T X c$, the squared covariance of the score and Y, the vector or matrix of concentrations. The loading, c_1, is the eigenvector of $X^T Y Y^T X$ and the score, t_1, the eigenvector of $X X^T Y Y^T$. The corresponding eigenvalue gives the squared covariance. As for PCR, PLS decomposes X as a bilinear form as in identity 2.4 and at the same time Y is decomposed as

$$Y = t_1 q_1^T + t_2 q_2^T + \ldots + t_r q_r^T + f,$$

where the q_is are found by regressing Y on the t_is. Subsequent loadings and scores are calculated iteratively by projecting X onto the subspace orthogonal to c_1 and Y onto the subspace orthogonal to the score t_1 and repeating the process replacing X and Y by the resulting matrices. As with PCR the process is continued until an appropriate number of regressors has been found. The number of regressors may be determined by cross validation.

Another description of the set of loadings used for PLS is given in Helland (1988). Helland proves that, provided that $s, Ss, S^2s, \ldots S^{m-1}s$ are linearly independent,

$$< c_1, c_2, \ldots, c_m >=< s, Ss, S^2s, \ldots, S^{m-1}s >$$

where $s = X^TY$, $S = X^TX$ and $< \ldots >$ denotes the span of the enclosed elements. $\{s, Ss, S^2s, \ldots, S^{m-1}s, \}$ is known as the Krylov basis for the loadings.

The scores for PLS are selected so that they are likely to contain information useful for predicting Y, so PLS does not suffer from the disadvantage that PCR does. In fact for our applications, PLS usually works better than PCR, producing more accurate predictions requiring fewer regressors.

For a detailed discussion of both PCR and PLS and the relationship between them see Stone and Brooks (1990). For an analysis of PLS see Helland (1988).

2.4.4 Cross-validation

Methods such as PCR and PLS where a decision has to be made about the number of scores to be used in the regression require two sets of samples, a calibration set to construct scores and a validation set to determine the optimum number of scores. An alternative to this which does not require a separate validation set is cross-validation. In cross-validation one sample is removed from the calibration set and the remaining samples used for calibration. The prediction error is calculated on the sample that was removed. The sample is then replaced. The procedure is repeated for each sample in the calibration set. The total squared error of prediction using the removed samples is calculated and this is used to assess the calibration. The process is repeated using different numbers of scores and the total squared errors of prediction are used to select the most appropriate number of scores. This can be done by plotting the errors and adding scores until the reduction in the total squared error of prediction is small.

2.4.5 Ridge regression

The technique of ridge regression was first introduced to deal with the problem of the matrix X^TX being singular or near-singular . In an attempt to remedy this Hoerl and Kennard (1970) suggested replacing X^TX by (X^TX+kI) in the expression for the least squares estimate of β to give the ridge estimate. Hoerl and Kennard show that the regression coefficient produced by ridge regression is the least squares estimate subject to the condition that its squared length is equal to a given constant; i.e. subject to the condition that β lies on the surface of a hypersphere of given radius. They prove that there exists a non-zero value of k for which the mean squared error between the true regression coefficient and the ridge estimator is a minimum so that if minimum mean squared error rather than least squares error is the criterion by which β is estimated, by selecting the appropriate value of k, ridge regression will provide a better solution.

Ridge regression can be generalised by replacing kI by a diagonal matrix, K, so that in the solution of the regression equation β is now constrained to lie on an ellipsoid. Goldstein and Smith (1974) showed that there always exists a better solution, ie. a smaller minimum mean squared error using generalised ridge regression.

An alterative justification for ridge regression is given by Lindley and Smith. In a Bayesian interpretation Lindley and Smith (1971) show that the ridge constant, k, can be estimated as the ratio of the variances of the model error and the prior for the regression coefficients, β_j, where in the prior, $\beta_j \sim N(0, \tau^2)$. Their result indicates that ridge regression is valid even if X is not ill-conditioned, identifying it as a form of Bayesian shrinkage. Since the analysis depends on the regression coefficients being exchangeable, the result also shows that ridge regression is inappropriate where this is not the case.

Shrinking the regression coefficients is also the motivation for Goldstein and Smith's approach (Goldstein and Smith (1974)). They suggest that since small

eigenvalues are known to inflate the least squares estimate of β the problem can be overcome by shrinking the estimate. They define a class of shrinkage functions which are dependent on the eigenvalues λ_i of X^TX and constants k_i for the regression coefficients, of which the simplest yields the Hoerl-Kennard ridge estimator.

Marquardt (1970) in a paper which compares the ridge estimate with an estimate produced using a generalised inverse, shows that the ridge estimator is equivalent to the least squares estimator for a dataset augmented by a set of orthogonal data whose reference values are all set to zero. This result can clearly be adapted to apply to the generalised ridge estimator. Marquardt shows that the generalised inverse solution to the regression equation is the best (in the sense that it minimises the sum of squares of residuals) least squares solution in the subspace spanned by the eigenvectors of X^TX which are assumed to have non-zero eigenvalues.

Another characterisation of the generalised inverse solution is as a generalised ridge estimator with $k_i = 0$ if $\lambda_i \neq 0$ and $k_i = \infty$ if λ_i is assumed to be zero (Goldstein and Smith (1974)). If we transform to the canonical form (Goldstein and Smith (1974)) then in terms of Lindley and Smith's model this is equivalent to placing vague priors on the γ_i, the transforms of the β_i, corresponding to non-zero eigenvalues, but setting equal to zero those γ_i whose eigenvalues are assumed to be zero.

In chapter 10 we show that the use of a repeatability file as a method of instrument standardisation is related to ridge regression.

2.5 Data

We used three different examples of NIR spectroscopic data. The first (NIR data) is from Wise's PLS Toolbox for Matlab (Wise and Gallagher (1998)) and consists of absorbances at 401 frequencies scanned at 2 nm. intervals in the range 800 to 1600 nm. for 30 samples of pseudo-gasoline data, measured

on two different NIR instruments. As well as absorbances, concentrations of five compounds present in the samples are given. The second dataset (CORN data) contains absorbances at 700 frequencies scanned at 2 nm. intervals in the range 1100 to 2498 nm. for 80 samples of corn measured on three different instruments together with four reference values, the concentrations of moisture, oil, protein and starch for each sample. This dataset is available as a MATLAB file on `www.eigenvector.com/Data/Corn`. The third set (BARLEY data) consisted of 85 samples of barley scanned on NIRSystems 6500 instruments at the same absorbances as for the CORN data, on 7 different instruments, together with protein content for each sample. For each data set, an arbitrarily chosen master instrument was calibrated using all the data for that instrument. Calibration was performed using PLS, the number of factors being determined by cross-validation. Graphs of the spectra for all of the samples on the master instruments for the three data sets are shown in figures 2.1, 2.2 and 2.3. Graphs of observed against predicted reference values are shown in figures 2.4, 2.5 and 2.6 for the NIR data, 2.7 and 2.8 for the CORN data and 2.9 for the BARLEY data.

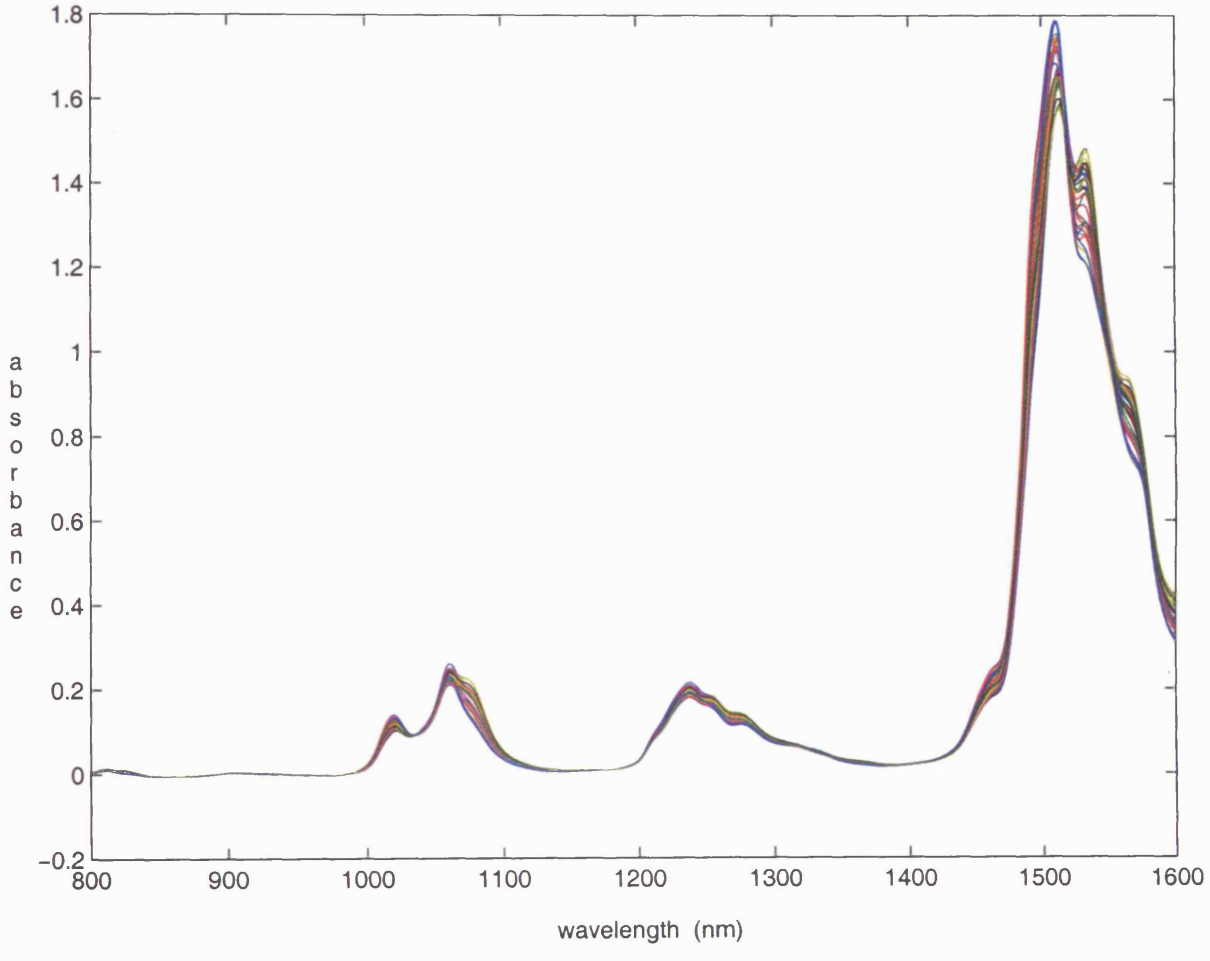


Figure 2.1: NIR data: graph showing spectra for 30 samples measured on the master instrument

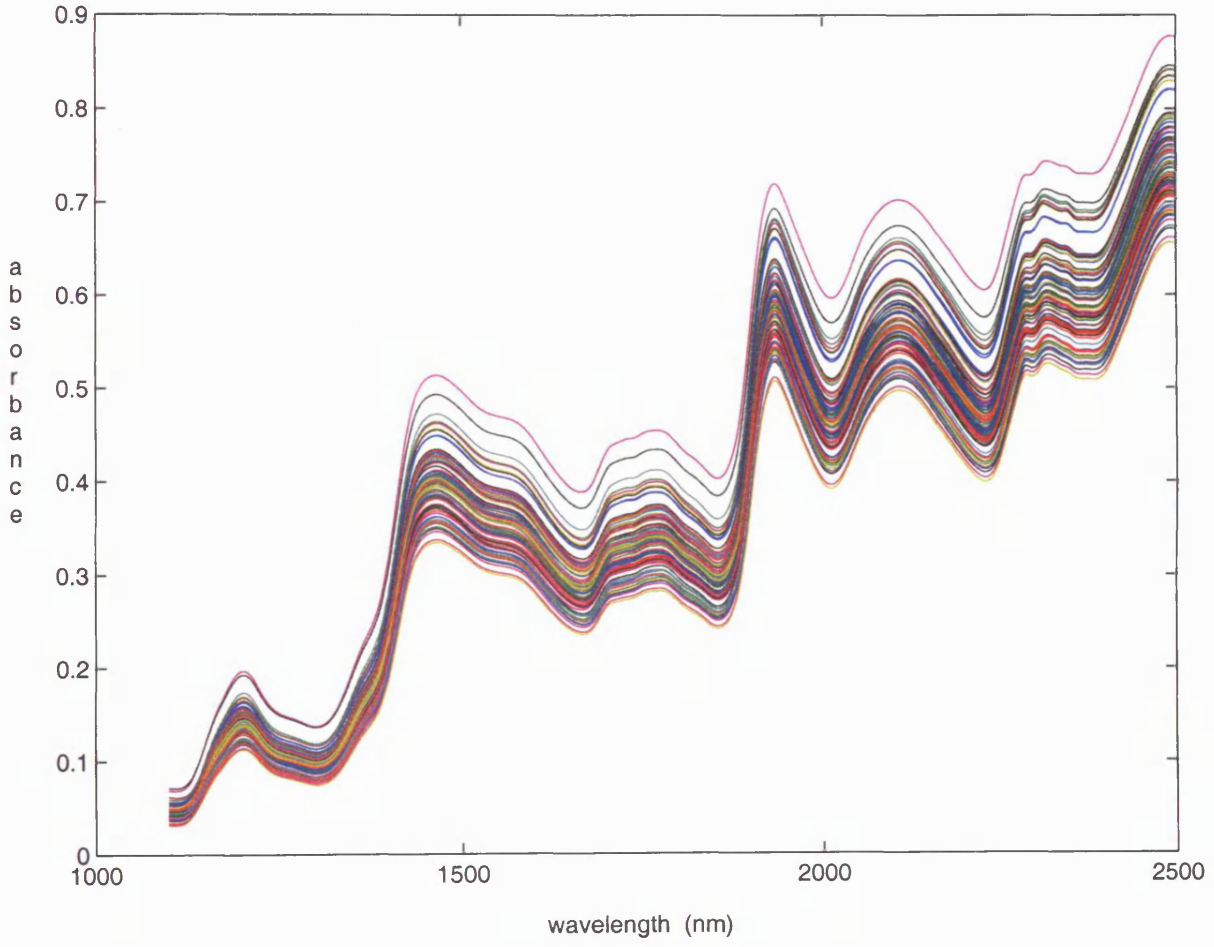


Figure 2.2: CORN data: graph showing spectra for 80 samples measured on the master instrument

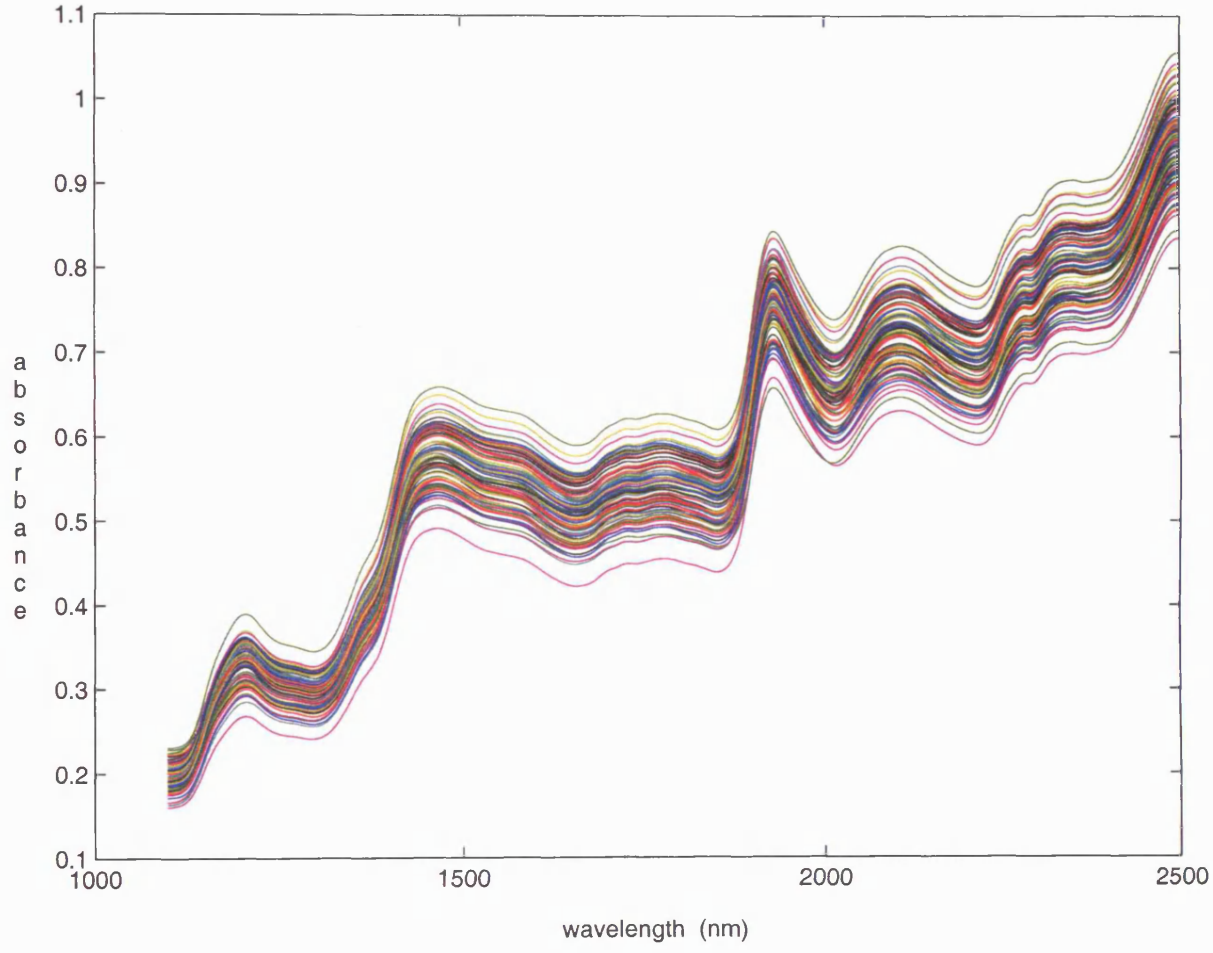
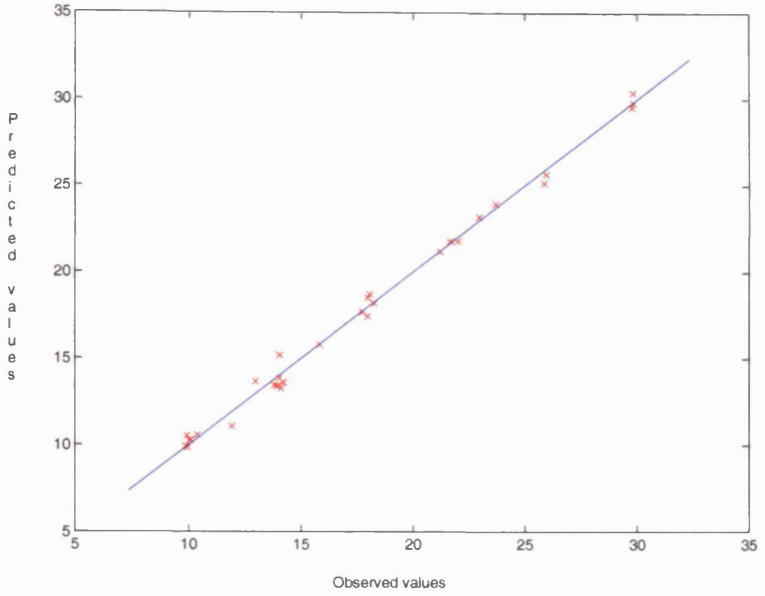
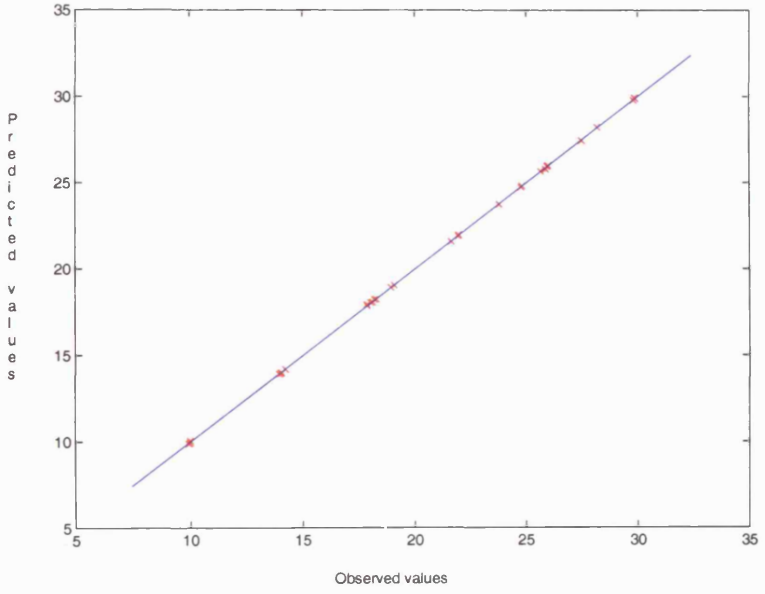


Figure 2.3: BARLEY data: graph showing spectra for 85 samples measured on the master instrument

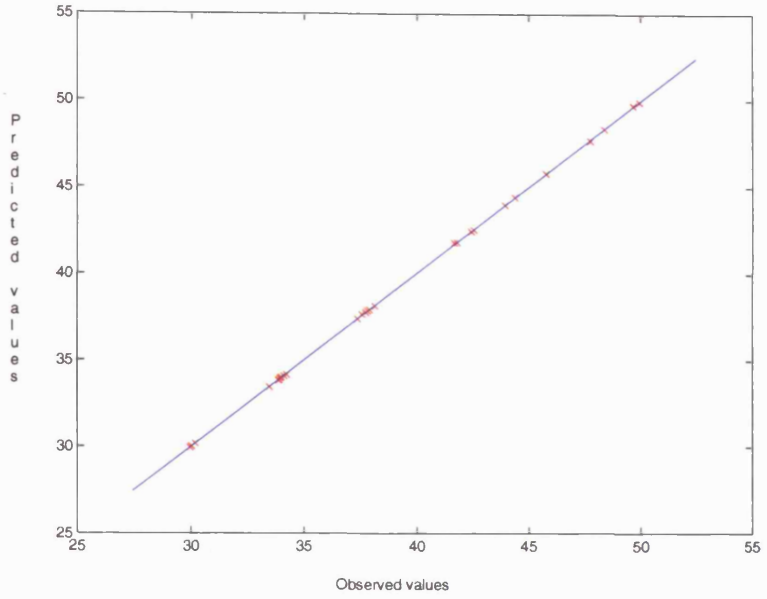


(a)

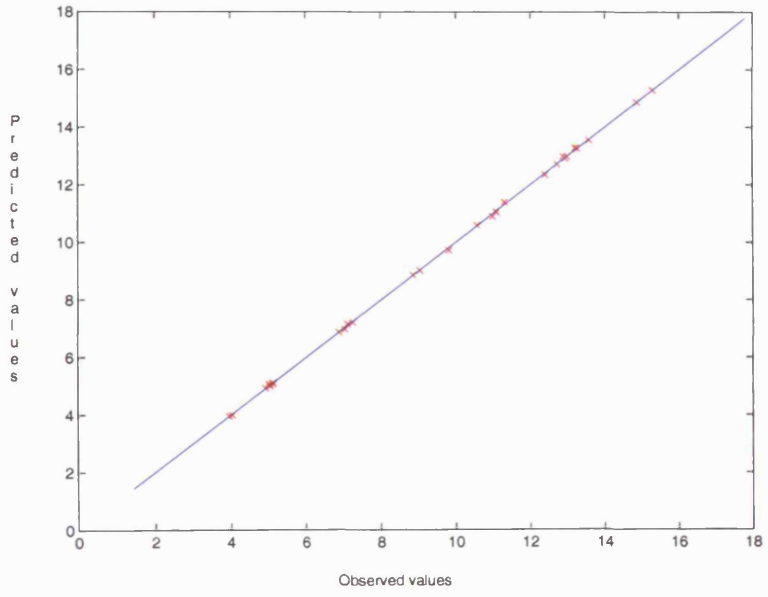


(b)

Figure 2.4: Graphs of observed against predicted reference values 1 and 2 for NIR data. 5 and 8 factors respectively used in calibration



(a)



(b)

Figure 2.5: Graphs of observed against predicted reference values 3 and 4 for NIR data. 8 factors used in each calibration

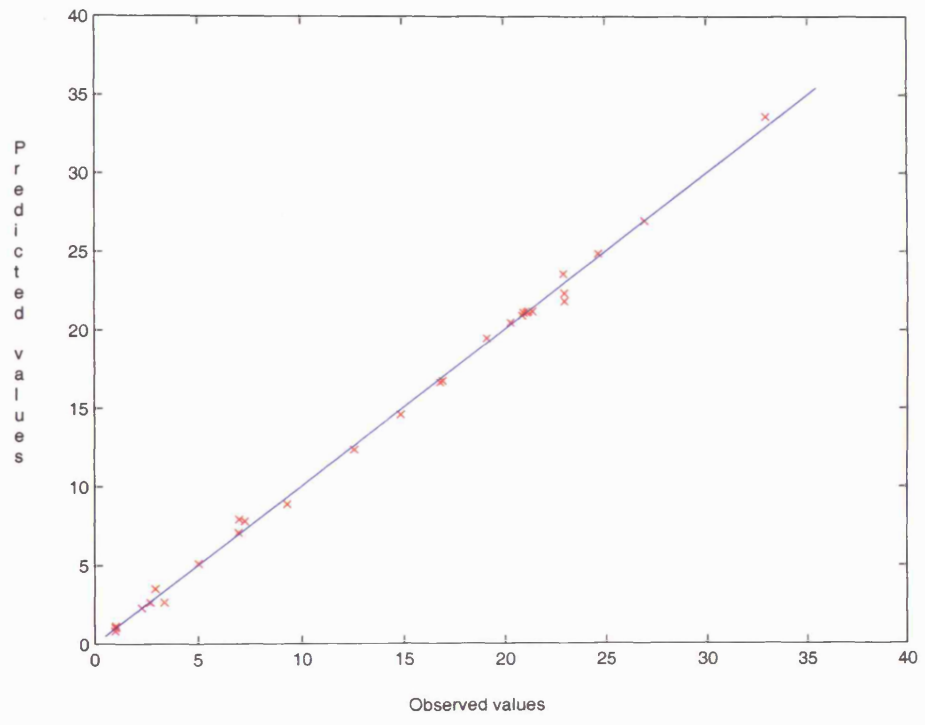
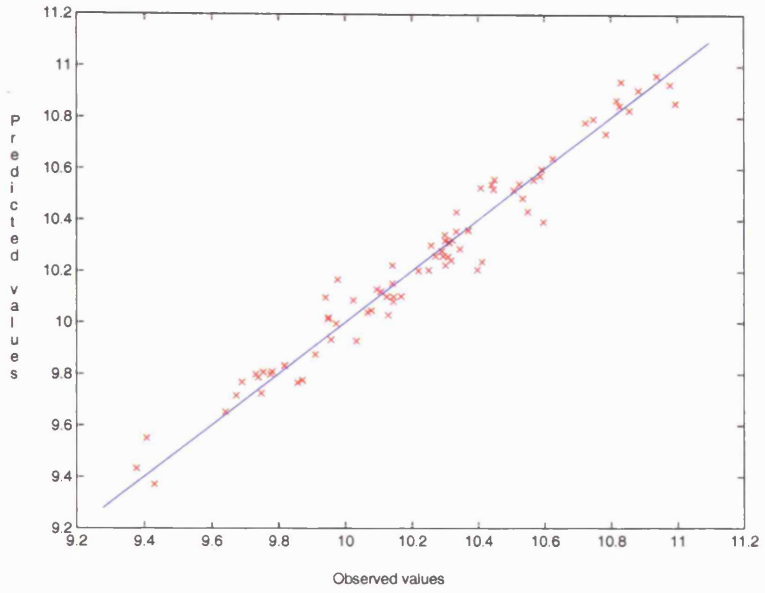
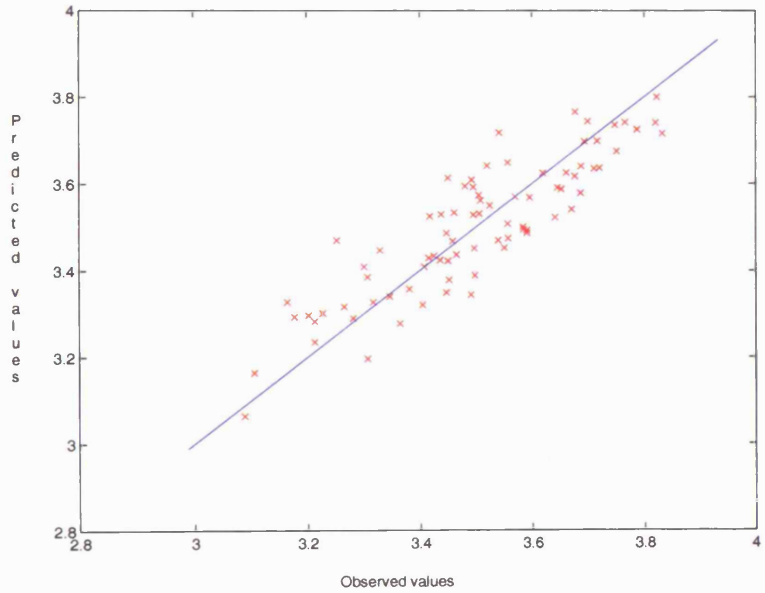


Figure 2.6: Graphs of observed against predicted reference value 5 for NIR data. 5 factors used in calibration

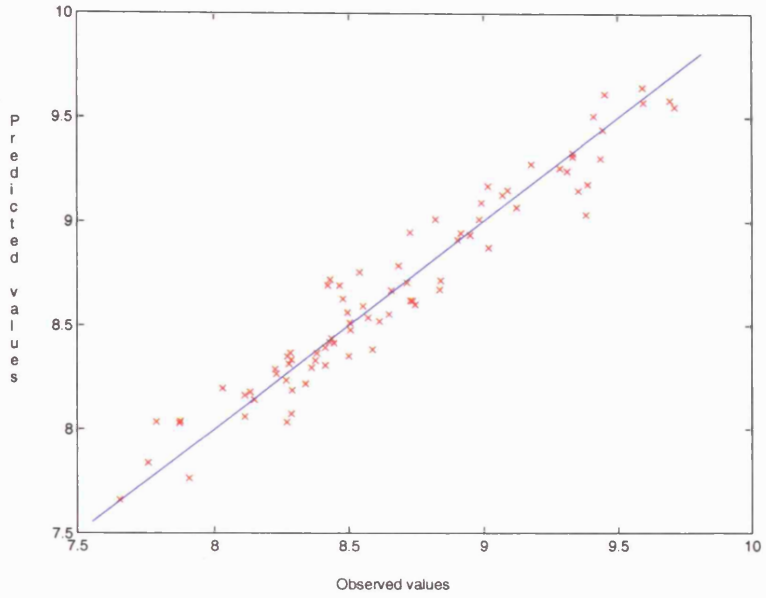


(a)

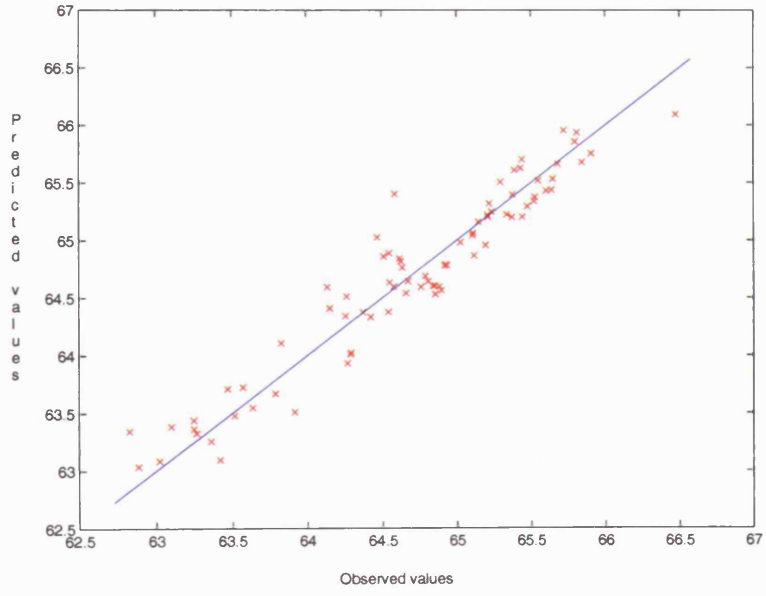


(b)

Figure 2.7: Graphs of observed against predicted reference values 1 and 2 for CORN data. 4 factors used for each calibration



(a)



(b)

Figure 2.8: Graphs of observed against predicted reference values 3 and 4 for CORN data. 6 and 7 factors respectively used for calibration

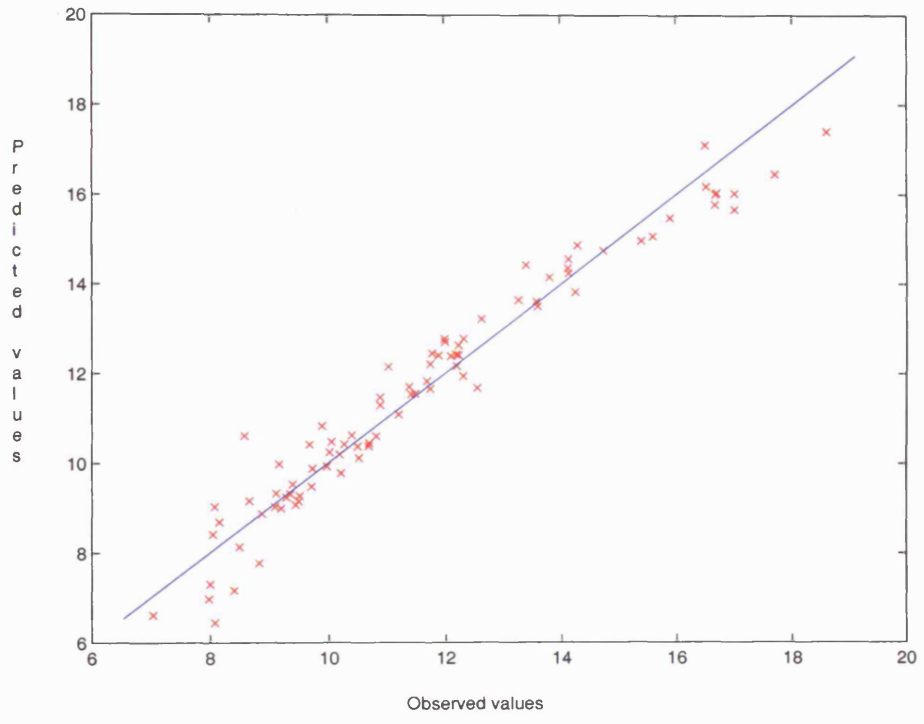


Figure 2.9: Graphs of observed against predicted reference values for BARLEY data. 11 factors used in calibration

Chapter 3

Standardisation

3.1 Introduction

Spectral responses of different instruments may vary for a variety of reasons. Any of the main components - light source, optical system or detector may respond differently in different instruments and even in the same instrument in differing atmospheric conditions or at different temperatures. A second but related problem is that the response of the same instrument may 'drift' over time or may alter after its has undergone some change due to servicing or repair. The calibration of a near-infrared spectrometer may involve several hundred samples, for which spectra and reference values are required, collected over a considerable period of time, and consequently the process is both costly and time-consuming. Calibration samples are chosen to reflect the types of materials for which a particular instrument is to be used, so may be perishable or not easily transportable. It is therefore important to be able to transfer the calibration to a second instrument or continue to use it after the instrument has been in service for some time without having to recalibrate. The process of adjustment to enable the original calibration to work on other instruments is known as instrument standardisation.

There are two main causes of differences between spectral responses on

different instruments. The first is wavelength, or horizontal, shift, where the instruments do not make measurements at precisely the same wavelengths. The second is absorbance (vertical) shift. Here absorbance measurement differs between instruments. Both differences can occur at the same time.

Much effort has been devoted to the problem of instrument standardisation. Accounts can be found in various reviews. Dean and Isaksson (1993a) give a concise overview, while Dean and Isaksson (1993b) provide a good assessment of the main methods. A more detailed account of the main methods is given in Dean and Kowalski (1996). de Noord (1994) and Bouveresse and Massart (1996) give well-constructed, detailed accounts of the main categories of standardisation methods and the most important methods used. More up to date accounts can be found in Fearn (2001) and Feudale et al. (2002), each of which includes a comprehensive list of references. Here we mention the main approaches to standardisation, with details of those methods that are relevant to our research.

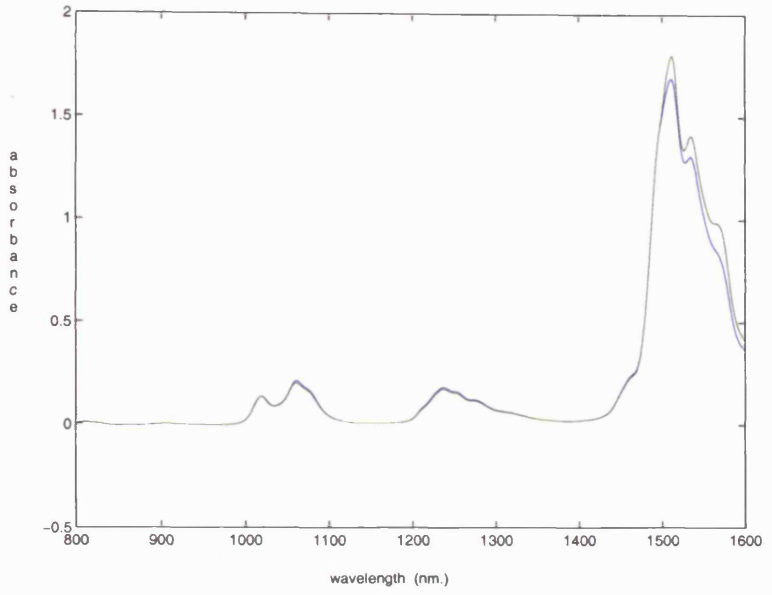
3.1.1 Examples

In figures 3.1 and 3.2 are shown, for each of the data sets introduced in chapter 2, spectra for a single sample measured on each of the different instruments in the set.

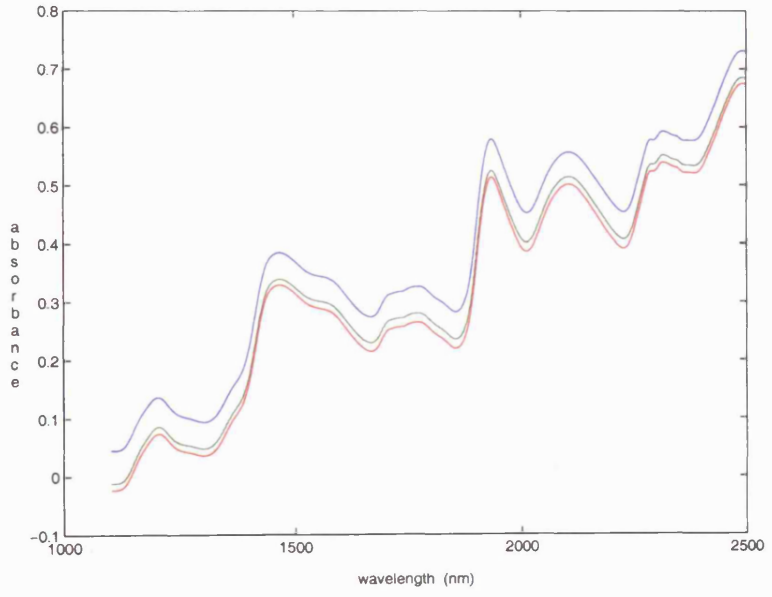
The master instrument was calibrated using PLS with the number of factors used determined using cross-validation, as described in the previous chapter. The root mean square error of calibration (RMSECV) was then calculated:-

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^c)^2}{n}}$$

where n is the number of samples in the calibration set, y_i, $i = 1, \ldots, n$ are the reference values for the n samples and y_i^c are the reference values estimated in the cross-validation. The same calibration equation was then used to estimate the predicted reference values on the remaining instruments in the set (the



(a)



(b)

Figure 3.1: Graphs showing the spectra of the same sample measured on each of the different instruments. a) NIR, b) CORN

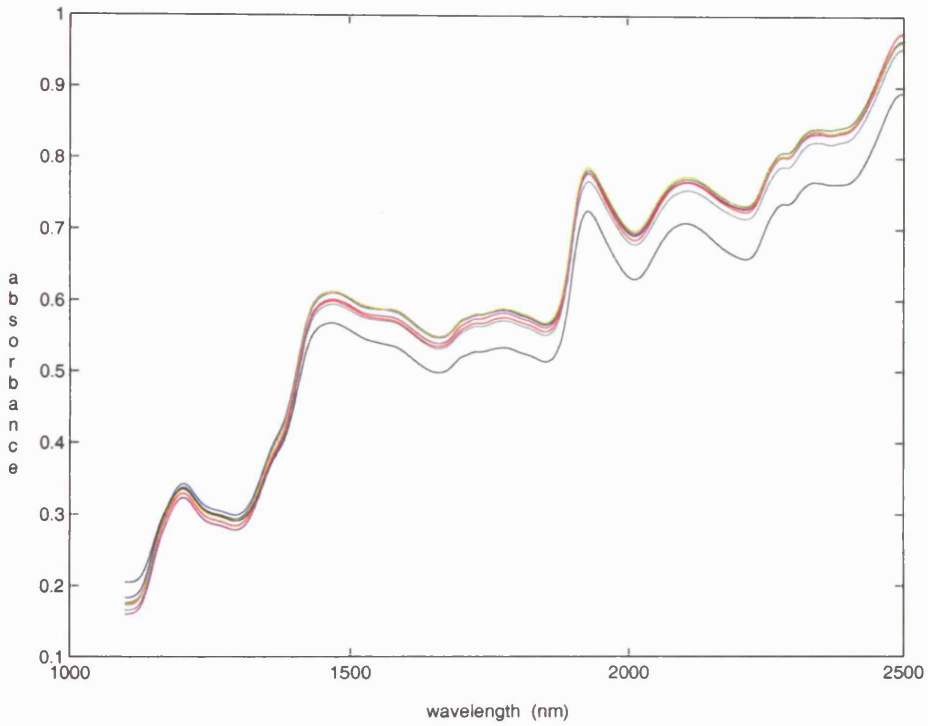


Figure 3.2: Graph showing the spectra of the same sample of BARLEY measured on each of the different instruments.

Table 3.1: NIR data

Reference value	1	2	3	4	5
no. of factors	5	8	8	8	5
RMSECV	0.6363	0.1185	0.0717	0.1268	0.5696
RMSEP	9.3995	1.2074	1.4869	2.8324	3.6518

Table 3.2: CORN data

Reference value	1	2	3	4
no. of factors	4	4	6	7
RMSECV	0.0862	0.0986	0.1677	0.2944
RMSEP (instrument 1)	1.4816	0.1672	0.7384	2.1927
RMSEP (instrument 2)	1.4821	0.1679	0.7385	2.1977

Table 3.3: BARLEY data. Master and 6 slave instruments.

RMSECV 11 factors	RMSEP					
Instrument no.	1	2	3	4	5	6
0.6412	0.9617	1.9962	0.7685	0.5116	1.1957	2.4209

slave instruments) and the root mean square error of prediction (RMSEP) for each of the slave instruments was calculated. For each of our three datasets we give, in tables 3.1, 3.2 and 3.3, the number of factors used in calibration, the root mean square errors of calibration (RMSECV) for each of the master instruments and the root mean square error of prediction (RMSEP) for the slave instruments, for each of the concentrations. Throughout this thesis we use these same calibration equations.

With the exception of two of the slave instruments for the barley data, the RMSEP on the slave instruments are so large compared with the RMSECV that the predicted values are clearly very inaccurate. These tables illustrate the problem that arises when a calibration is transferred between instruments

and hence the need for standardisation if the calibration of each instrument separately is to be avoided. The same conclusion can be drawn from the graphs in figures 3.1(a), 3.1(b) and 3.2. The differences between the slave and master for the NIR spectra are greatest in the range 1500 - 1600 nm. where most of the information lies. For the corn data, the difference between slave spectra and the master spectrum involves a vertical shift, possibly due to the effect of light scatter.

3.1.2 Notation

Throughout this thesis we denote by $X_m(i,j)$ the absorbance on the ith sample at the jth wavelength on the master instrument and by $X_s(i,j)$ the corresponding absorbances on a slave instrument. y_i denotes a response on the ith sample.

3.2 Standardisation sets

Most standardisation methods rely on a standardisation set, a set of carefully chosen samples for which spectra are available on both the master and slave instruments. Some methods also require reference values for the samples.

The number of samples needed to standardise an instrument depends on the complexity of the differences between the master and slave instrumental responses and the standardisation method used. Where there is only a vertical shift, (absorbance difference) standardisation with a single sample may produce acceptable results for the simplest methods. When a wavelength shift is present and a complex method is used, more samples may be needed. For practical reasons the standardisation set should be as small as possible. Clearly if it is as large as the calibration set recalibration would be possible and preferable. We have used standardisation sets of sizes varying between one and ten samples.

Various types of standardisation set have been produced for use commercially. Shenk and Westerhaus (1989) produced a set of thirty dried agronomic products in sealed containers to be used with their patented method. The spectra for these samples were very similar to samples of corn and grass which were used in prediction. In a attempt to provide a general purpose standardisation set which could be used to standardise spectrometers used for a wide variety products, Dardenne and Biston (1991) made a set of twelve organic and inorganic compounds, designed to cover a wide range of optical densities (i.e. absorbances). A third set of six substances with almost flat spectra, but covering a wide range of optical densities, has also been tested. The attraction of these samples is that they are easily transportable and should not deteriorate quickly and can be used to standardise instruments calibrated for a variety of products. However results from research by Bouveresse et al. (1994), Bouveresse et al. (1995) suggest that the standardisation samples that work best are those that are similar to the substances for which the instrument is eventually to be used.

Attempts to produce general purpose standardisation sets have not usually proved successful. This is to be expected since most standardisation methods assume a linear relationship between spectra on slave and master instruments. This is an approximation and will only be valid over a narrow range of optical densities. Over a wider range a linear model will be inadequate even within the range of the standardisation samples and beyond this results are likely to be unreliable.

In experimental situations it is common to use a subset of the calibration set for standardisation. Two methods of subset selection are frequently used. The first is due to Wang et al. (1991) and is based on selecting samples with high leverages. A second method, due to Kennard and Stone (1969), selects samples whose Euclidean distances from each other are maximised.

Wang's method uses the mean-centred matrix of spectra and first selects

the sample whose sum of squares, $\sum_j (X(i,j) - \bar{X}(j))^2$, where $\bar{X}$ is the mean of the samples, is greatest. The mean-centred matrix of spectra is then projected onto the subspace orthogonal to the selected sample. The process is repeated using the resulting matrix and continues until the required number of samples has been selected.

In Kennard and Stone's method, the first two samples are selected by calculating the Euclidean distance $\sqrt{(\sum_j (X(i,j) - X(k,j))^2)}$ between samples i and k for each pair of samples, and selecting the pair for which this distance is greatest. Further samples are added by calculating the distance of each sample not already selected from each of the selected samples. The sample whose minimum distance from any of the already selected samples is greatest is added to the set. The process is continued until the required number of samples has been selected.

Throughout this thesis we have used the method due to Wang et al. (1991) to select standardisation sets.

3.3 Standardisation techniques

Standardisation techniques fall roughly into three groups categorised by the point in the process of calibration and prediction at which the standardisation occurs. Firstly there are those methods that aim to produce a calibration that is robust to different instruments or conditions. Secondly there are methods that adjust the spectra from the slave instruments so that they match the corresponding spectra of the master instrument, enabling the same calibration to be used on both. Finally there are methods that adjust the concentrations that are predicted as a result of applying the master calibration to the slave instrument so that they match those that have been predicted by the master instrument or known reference values.

3.3.1 Pretreatments and robust calibrations

Pretreatmants are applied to spectra before calibration to remove certain differences between spectra. By calibrating after the pretreatment has been applied it is hoped that a calibration robust to those differences will be found.

A major cause of differences between spectra is light scatter. Light scatter depends on the refraction of light within samples and varies with particle size within samples and the refractive indices between particles and their surroundings. It is generally assumed to cause either a constant additive or multiplicative difference (or both) between spectra. Methods aimed at removing this type of difference are known as pre-treatments. Among the simplest of these are derivative treatments. If a first derivative treatment is applied it has the effect of removing differences due to a constant vertical shift. A second derivative treatment will also remove constant slope differences.

These treatments are frequently used because they are easy to apply and often effective. The only complication is that because spectra are stored as discrete points derivatives are not defined. There are various methods for estimating the derivative in these circumstances, the simplest of which is to define the derivative at the nth point as the difference between consecutive points:- $x_n - x_{n-1}$. A variation on this which is quite commonly used is to define the derivative at the nth point as $\sum_{n-r}^{n-s} x_i - \sum_{n+s+1}^{n+r+1} x_i$, $(r \geq s)$. Strictly speaking this expression should be divided by $r-s+1$, but often this is omitted. An example of this can be found in Shenk and Westerhaus (1991) where the method is used with $r = 3$ and $s = 0$. A more sophisticated approach due to Savitzky and Golay (1964) is to select a window of points surrounding the nth point and fit a polynomial to the points in the window. The derivative of the spectrum at the nth point is then defined as the derivative of the polynomial at that point. Wang and Kowalski (1992) fit a cubic to a 9-point window to estimate the derivative. All their results are based on first derivative spectra.

A more complicated pre-treatment, multiplicative signal correction, (Geladi

et al. (1985)), removes additive and multiplicative shifts. Each spectrum is modelled as differing from the mean of all the spectra in the set by an additive and a multiplicative constant and a term containing chemical information:-

$$x_i = a_i + b_i\bar{x} + \epsilon_i.$$

Here, x_i represents the vector of the spectrum of sample i, $\bar{x}$, is the mean of the set of spectra and ϵ_i is an error vector. For each sample the two constants are found by regressing the spectra for that sample on the mean spectrum, over all wavelengths. Each spectrum is then corrected to eliminate the additive and multiplicative differences.

Orthogonal signal correction (OSC) aims at filtering out spectral information that is irrelevant to the calibration. This is done by subtracting from X the matrix of spectral variation in X that is orthogonal to Y, the matrix of concentrations. The method was originally suggested by Wold et al. (1998). In their method, the first principal component Xc of X, where c is the loading, is found. Xc is then projected onto the subspace orthogonal to Y, giving $t_1 = (I - Y(Y^TY)^{-1}Y^T)Xc$. The score, p_1 and the loading t_1 are re-estimated by a PLS-type iteration. Once convergence is reached, the method proceeds as for PLS, subtracting $t_1p_1^T$ from X and repeating the process using $X - t_1p_1^T$ instead of X. The resulting matrix, $X - \sum t_ip_i^T$ is then used in calibration in place of X.

An alternative is suggested by Fearn (2000). His method does precisely what Wold's aims to do: maximise the variance of the selected scores for X subject to the constraint that they are orthogonal to Y.

There is an obvious attraction to removing noise from X before calibrating and Wold et al. (1998) report reduced RMSEP on test data used to predict viscosity for cellulose sheets, as well as fewer factors in the calibration equation. Fearn (2000), on the other hand, found that performance was unchanged when his method was used for the prediction of protein content of wheat samples, and that the reduction in factors required for calibration was exactly matched

by the number of factors removed by OSC.

Details of these and other pre-processing methods can be found in Naes et al. (2002).

The preprocessing methods described above are designed to correct absorbance shifts. Mark and Workman Jr (1988) developed a method which aimed to produce calibrations which were robust to wavelength shifts. It involves selecting wavelengths for use in calibration for which regression coefficients in the calibration equation remain constant when the wavelength is altered slightly. This has the disadvantage that it may restrict the choice of wavelengths to those where there is less information and which therefore give less satisfactory calibrations.

Calibrations robust to different instruments can be produced by augmenting the calibration set by including spectra with their reference values from those instruments, an example of this can be found in Hardy et al. (1996). Although the calibration may perform less well on the master instrument, it is hoped that it will perform adequately on all instruments included. A development of this method known as the repeatability file method has been proposed by Westerhaus (1991) and tested by Shenk and Westerhaus (1991) and Tillmann et al. (2000). This method is described in detail in Chapter 10 of this thesis where we compare it to the method known as transfer by orthogonal projection (TOP) which we have developed and tested.

3.3.2 Methods that adjust the spectra

Several standardisation methods have been devised that adjust the slave spectra to match the master spectra as closely as possible. The calibration developed for the master spectra can then be used on the adjusted slave spectra. Standardisation methods that adjust spectra have proved successful in terms of minimising prediction errors, especially in correcting for a wavelength shift. The Shenk-Westerhaus patented method (S-W) is widely used commercially,

as is Piecewise Direct Standardisation (PDS) which the literature (Wang et al. (1991)) suggests works better. Both methods are shown to be superior to methods that fall into the other two categories, i.e. methods that produce robust calibrations and those that adjust concentrations. (See for example Wang et al. (1991) and Bouveresse et al. (1996)).

Shenk-Westerhaus patented method

The Shenk-Westerhaus patented method (Shenk and Westerhaus (1989)) corrects for the two types of error, horizontal and vertical shift, separately. The method requires spectra for the same samples from both the master and the slave instruments, but no concentrations are needed. The wavelength shift is corrected first. This is done wavelength by wavelength using spectra to which a first derivative treatment has been applied. For each wavelength, j, the absorbances on the master instrument are compared with those on the slave instrument for a window $[j-k, j+l]$ of neighbouring wavelengths, $X_s(i, j-k:j+l)$. Correlation coefficients

$$C(j,r) = \frac{\sum_i \dot{X}_m(i,j)\dot{X}_s(i,r)}{\sqrt{\sum_i \dot{X}_m(i,j)^2 \sum \dot{X}_s(i,r)^2}}, r = j-k, \ldots, j+l$$

are calculated. Here $\dot{X}$ represents X mean-centred over samples and the index, i, ranges over the samples in the standardisation set. A quadratic is fitted to $C(j,r)$ and its maximum value within the window found. The point at which the correlation is maximum is assumed to be the wavelength on the slave instrument corresponding to the wavelength j on the master instrument. To smooth the wavelength shifts, a quadratic model is fitted over the whole wavelength range to the shifts found in the previous step and this model is used to define the wavelength shifts between master and slave. Once the wavelength shift is found, the absorbance of the corrected wavelength on the slave instrument is estimated by linear interpolation on the absorbance measurements at adjacent wavelengths. Finally the absorbance shift is corrected using linear

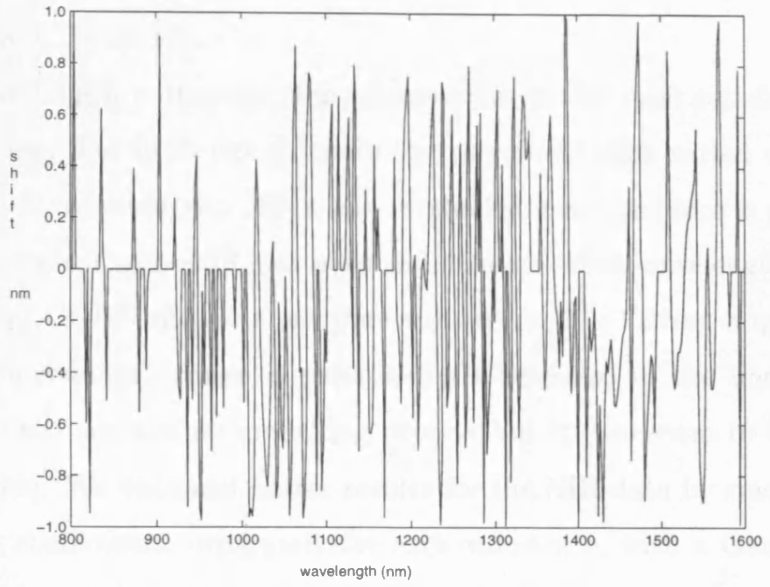
regression of slave absorbances on the master absorbances for the standardisation set, again treating each wavelength separately. The method is described in detail in Bouveresse et al. (1994) .

The Shenk -Westerhaus method was tested on three similar instruments using grass, corn and rapeseed samples by Bouveresse et al. (1994). As mentioned previously, they achieved good results only when the standardisation set consisted of materials which were similar to the samples for which predictions were to be made. In an attempt to compensate for the difference between standardisation samples and prediction samples, Bouveresse et al. (1995) used a technique known as locally weighted regression (LWR). This technique involves, for each wavelength, weighting the standardisation samples, giving greatest weight to the samples whose spectra are closest to the spectra for which predictions are to be made. LWR is applied at each wavelength separately and involves first calculating, $\delta(j)$, the total squared Euclidean distance of standardisation sample j on the slave instrument from all of the prediction samples:-

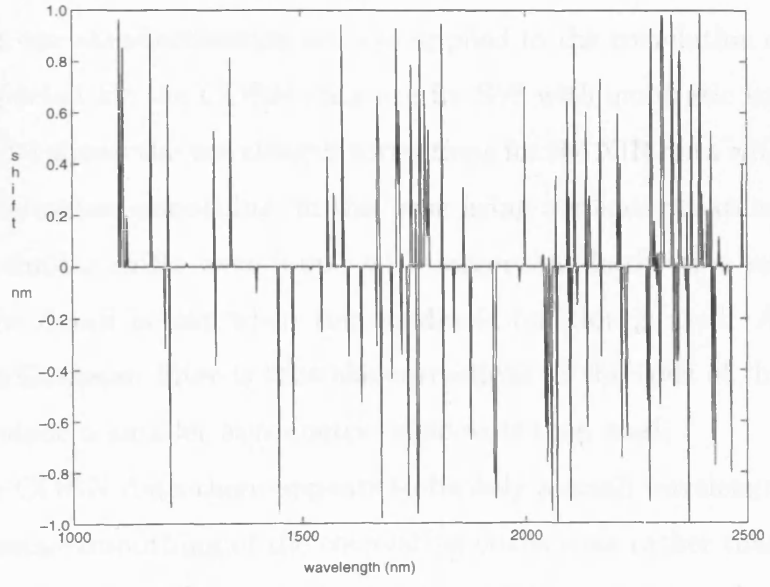
$$\delta(j) = \sum_i (X_s(j,k) - X_s(i,k))^2.$$

Here i indexes prediction samples and k is a fixed wavelength. The standardisation samples for which this distance is large are discarded, while the rest of the standardisation samples from both the master and slave instrument are weighted using weights between 0 and 1, giving the largest weights to those samples for which $\delta(j)$ is smallest. Weighted master absorbances are then regressed on weighted slave absorbances as usual. Using LWR resulted in improved results where the standardisation set consisted of samples which were very different from the prediction samples. However, the best results were still obtained by using standardisation samples that were similar to the prediction samples.

We programmed the Shenk-Westerhaus standardisation method in Matlab, following the description in Bouveresse et al. (1994) and applied the method



(a)



(b)

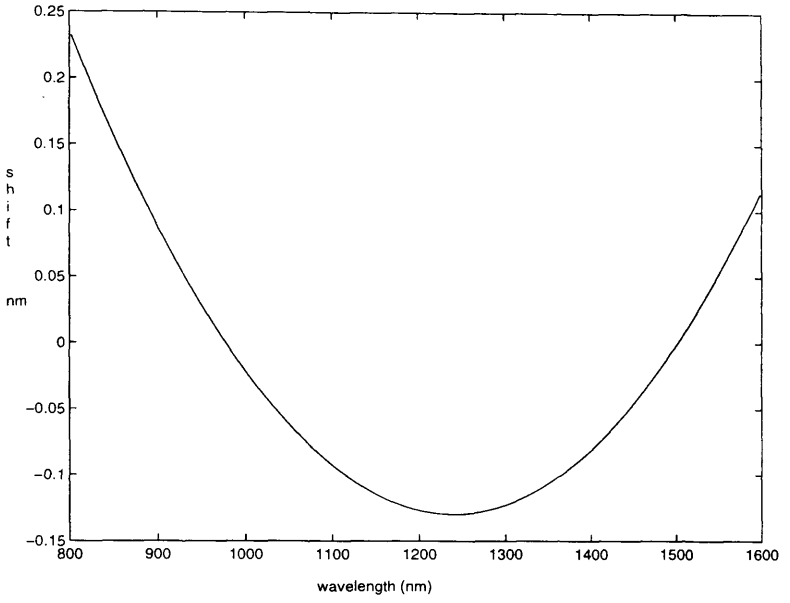
Figure 3.3: Unsmoothed wavelength shift for (a) NIR and (b) CORN data.

to the two datasets, NIR and CORN, that appeared to have wavelength shifts. A wavelength window of size 3 was used for both datasets and standardisation sets of sizes 5, 8 and 10.

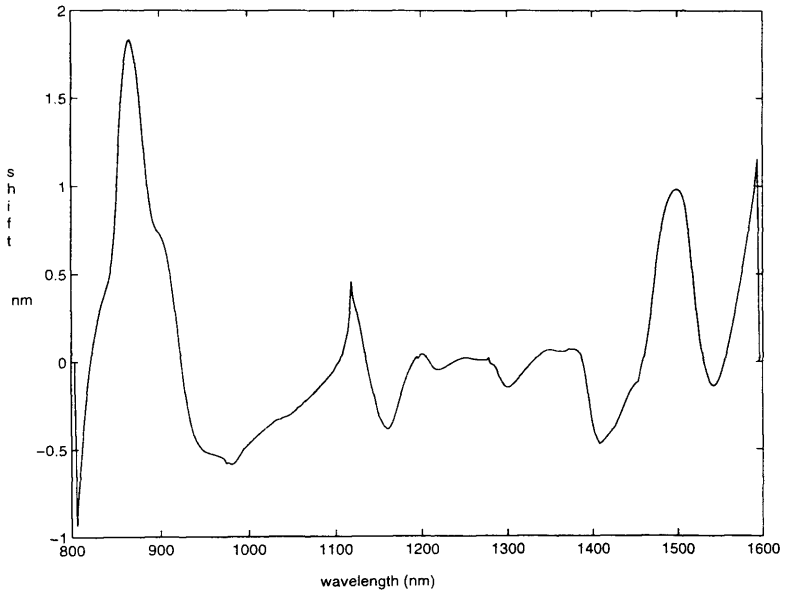
The wavelength correction step appeared to be the least satisfactory part of the process. For both our datasets the calculated shift varied wildly from wavelength to wavelength. Often the method did not produce a clear result and in this case a zero shift was assumed. Unsmoothed wavelength shifts for the NIR and CORN datasets are shown in figure 3.3. Fitting a quadratic to the entire wavelength range is presumably a response to the unsatisfactory outcome of the correlation modelling process but it does seem to be an over-simplification. We obtained better results for the NIR data by smoothing the correlation coefficients, separately for each value of r, with a Gaussian filter and a smaller window, before the wavelength shift was determined. For the NIR data as well as fitting a quadratic to the wavelength shifts as in the SW method, a Gaussian smoothing filter with a window of 41 selected to minimise RMSEP on the standardisation set was applied to the correlation coefficients. Results reported for the CORN data are for SW with quadratic smoothing.

Figure 3.4 shows the wavelength corrections for the NIR data with quadratic and with Gaussian smoothing, in this case using a standardisation set of size 5, though similar shifts were found with larger standardisation sets. Clearly much of the detail is lost when the quadratic function is used. A disadvantage of the Gaussian filter is that the corrections at the ends of the range are unreliable since a smaller asymmetric window is then used.

For the CORN data there appears to be only a small wavelength shift and using Gaussian smoothing of the correlation coefficients rather than quadratic smoothing made no difference. Again the correlation process does not work well. Quadratic smoothing or Gaussian smoothing using a very wide window suggests a small, possibly negligible, shift. Figure 3.5 shows the wavelength correction using quadratic smoothing.



(a)



(b)

Figure 3.4: Graphs showing the wavelength adjustments for NIR a) quadratic smoothing b) Gaussian smoothing of the correlation coefficients

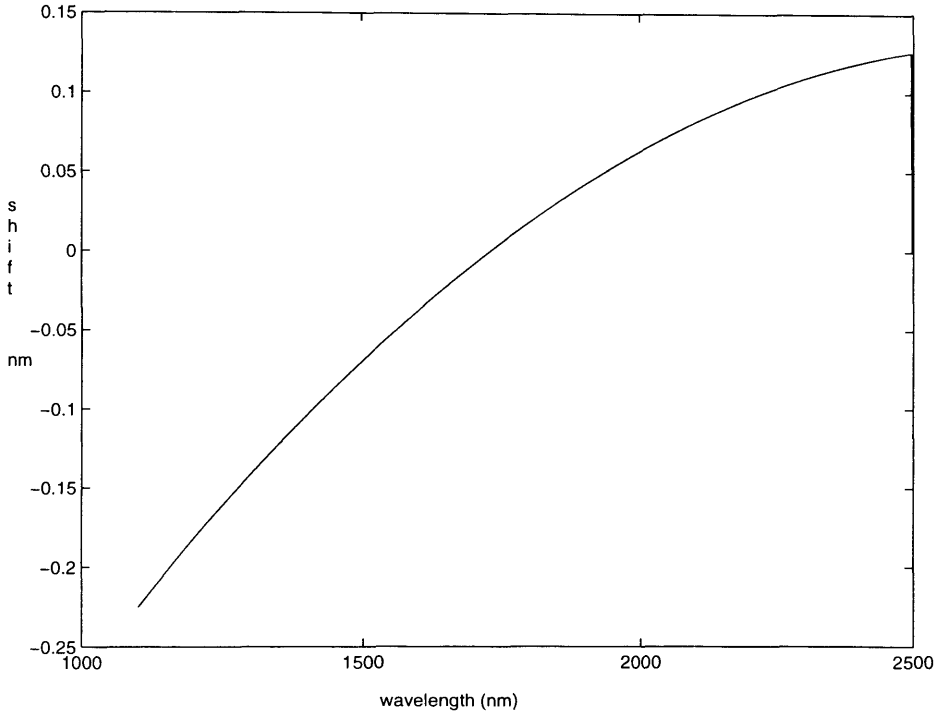


Figure 3.5: Graph of wavelength shift with quadratic smoothing for CORN data

The RMSEP calculated after standardisation was based only on those samples not used in the standardisation process. We also give, for the purposes of comparison, the RMSEP based on all the samples available. These results are given in tables 3.4 and 3.5. By comparing the RMSEP after using SW with the RMSEP on the unadjusted slave it can be seen that even with a small standardisation set, SW works well. For both datasets, increasing the size of the standardisation set improved perfomance and for the NIR data, Gaussian instead of quadratic smoothing of the wavelength shift function also improved performance.

Figure 3.6(a) shows the regression coefficients for the NIR data. It can be seen that the intercept is fairly close to zero, while the gradient is fairly close to 1 indicating that only small adjustments were necessary. The sharp peaks in the gradient graph occur where there is virtually no information in the spectra and so are unlikely to affect predictions. Figure 3.6(b), shows the regression coefficients for the CORN data. These have intercepts close to zero and gradients of approximately 1.2. The CORN data requires a smaller wavelength correction, but a larger absorbance correction than the NIR data.

Direct Standardisation

The method of direct standardisation (DS) was proposed by Wang et al. (1991) and is based on the assumption that there exists a relationship

$$X_m = X_s F \tag{3.1}$$

where F is a $w \times w$ matrix. It is assumed here that X_s and X_m have been mean-centred to avoid the need for an additive constant. F is determined using the known spectra on a standardisation set and the relationship

$$F = X_s^- X_m \tag{3.2}$$

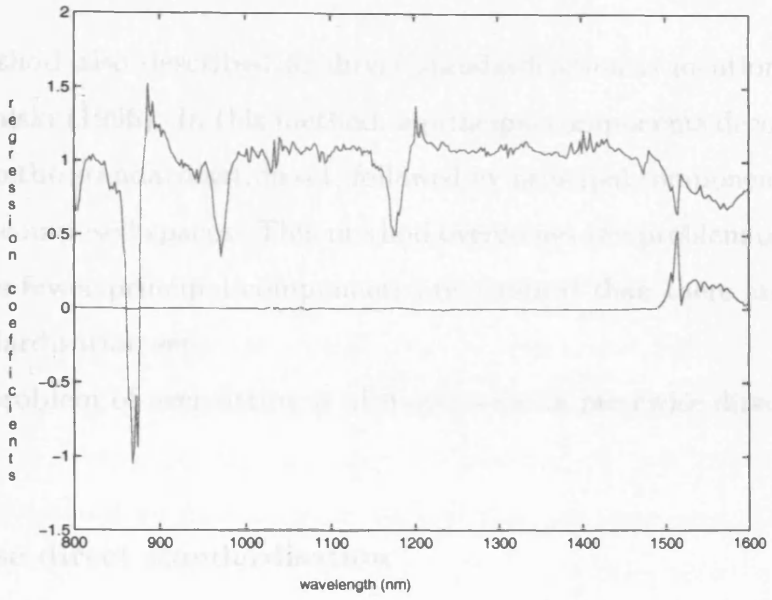
Here X_s^- denotes a generalised inverse of X_s. Since $n << w$ there is bound to be over-fitting since a perfect fit can be found for the standardisation set,

Table 3.4: NIR Data before and after applying SW, RMSEC and RMSEP for dataset omitting standardisation samples (25, 22, 20 spectra) and for entire dataset (30 spectra).

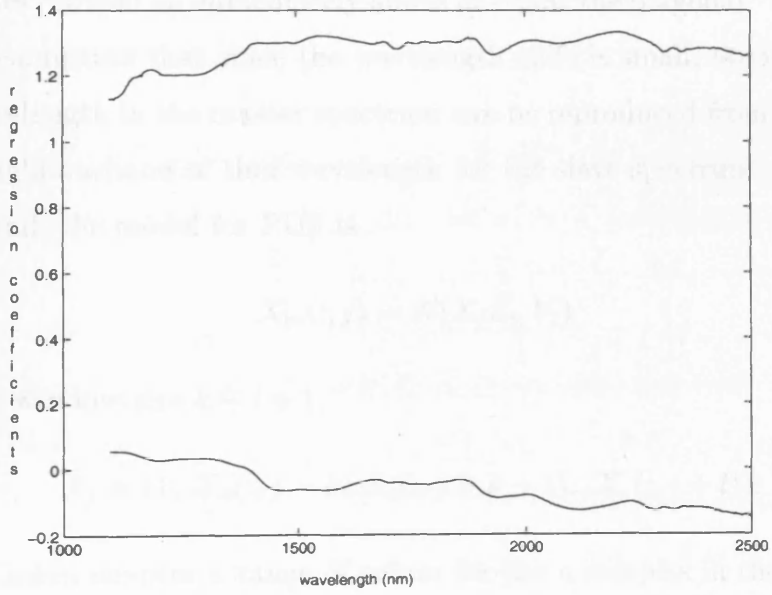
Reference value	1	2	3	4	5
Master	0.6363	0.1185	0.0717	0.1268	0.5696
Slave	9.3995	1.2074	1.4869	2.8324	3.6518
SW with quadratic smoothing: Standardisation set size 5					
25 spectra	2.3055	0.5495	0.1396	0.2732	1.7352
30 spectra	2.1381	0.5068	0.1319	0.2516	1.6010
SW with Gaussian smoothing: Standardisation set size 5					
25 spectra	0.7537	0.4406	0.1873	0.5255	0.5925
30 spectra	0.7679	0.4460	0.1798	0.5140	0.6412
SW with quadratic smoothing: Standardisation set size 8					
22 spectra	1.3254	0.3722	0.0782	0.1055	1.0467
30 spectra	1.3456	0.4075	0.0775	0.1214	1.0673
SW with Gaussian smoothing: Standardisation set size 8					
22 spectra	0.7464	0.4850	0.1193	0.3798	0.6103
30 spectra	0.8051	0.4724	0.1198	0.3838	0.6640
SW with quadratic smoothing: Standardisation set size 10					
20 spectra	1.4851	0.3878	0.1119	0.1486	1.1547
30 spectra	1.4623	0.4099	0.1056	0.1553	1.1422
SW with Gaussian smoothing: Standardisation set size 10					
20 spectra	0.6571	0.2420	0.1729	0.2534	0.5604
30 spectra	0.7503	0.2738	0.1721	0.2627	0.6208

Table 3.5: CORN Data before and after applying SW. RMSEC and RMSEP for dataset omitting standardisation samples (75, 72, 70 spectra) and for entire dataset (80 spectra)

Reference value	1	2	3	4
Master	0.0862	0.0986	0.1677	0.2944
Slave	1.4816	0.1672	0.7384	2.1927
SW: Standardisation set size 5				
75 spectra	0.4909	0.1645	0.3170	0.5131
80 spectra	0.4767	0.1629	0.3135	0.5093
SW: Standardisation set size 8				
72 spectra	0.3694	0.1134	0.1744	0.3924
80 spectra	0.3540	0.1115	0.1725	0.3779
SW: Standardisation set size 10				
70 spectra	0.3014	0.1080	0.1696	0.3575
80 spectra	0.2993	0.1068	0.1654	0.3492



(a)



(b)

Figure 3.6: Graphs of regression coefficients for NIR and CORN data using SW.

but the same regression coefficients are unlikely to provide a good fit for other data.

A method also described as direct standardisation is mentioned by Dean and Kowalski (1996). In this method, a principal components decomposition is applied to the standardisation set, followed by principal components regression on the decomposed spaces. This method overcomes the problem of over-fitting as long as fewer principal components are retained than there are samples in the standardisation set.

The problem of over-fitting is also overcome in piecewise direct standardisation.

Piecewise direct standardisation

The same model is used in piecewise direct standardisation (PDS) as in direct standardisation but the matrix F is assumed to be highly structured with non-zero entries only on or immediately above or below the diagonal. This is based on the assumption that since the wavelength shift is small, absorbances at a given wavelength in the master spectrum can be reproduced from absorbances in the neighbourhood of that wavelength for the slave spectrum.

In detail, the model for PDS is

$$X_m(:, j) \sim N(X_j \beta_j, V_j)$$

where for window-size $k + l + 1$,

$$X_j = (1_n \, X_s(:, j-k) \, X_s(:, j-k+1) ... X_s(:, j+l)).$$

Here the colon denotes a range of values for the n samples in the standardisation set. β_j is a vector of regression coefficients of length $k + l + 2$ and V_j is an $n \times n$ covariance matrix. Usually, $k = l$ to give a symmetric window but an asymmetric window might be appropriate if, for example, the wavelength shift were known to be positive.

The window-size must be determined by experiment. Using simulated data with horizontal shift varying between +2 and -2 channels Wang et al. (1991) found the predictions for one analyte, using PDS only began to deteriorate when k was greater than 4. They attributed this to over-fitting. With a second analyte $k = 7$ gave the best results. This was possibly due to a non-linear response change for which the additional regressors compensated. These results indicate one of the drawbacks of PDS; that, because of the correlation between spectra at adjacent wavelengths it does not correct the actual error but compensates for it, resulting in over-fitting. If the difference between the slave and master spectra is the result of a wavelength shift and an absorbance shift, then as well as the intercept, only the central regression coefficient, the one corresponding to $X_s(:, j)$, and regression coefficients either to the right or to the left of the central coefficient, but not both, should be non-zero. Which of the side coefficients is zero depends on the size and sign of the wavelength shift. In applying PDS we found that all coefficients were non-zero and that it was not possible to discover the direction of the shift from the regression coefficients.

The size of the standardisation set needed to obtain acceptable results depends on the complexity of the difference between master and slave spectra. Wang et al. (1991) selected a standardisation set consisting of three samples from a set of gasoline samples used to calibrate the master instrument. Good predictions were obtained after the slave instrument had been standardised using just these three samples. Wang and Kowalski (1992) in a separate study again found that standardisation sets of three samples enabled them to reach acceptable results using PDS with mean-centred data and no additive term, using the NIR and CORN datasets we have used. Bouveresse et al. (1996) in a paper comparing PDS, with and without an additive term, with the slope/bias correction method (see section 3.3.3) used standardisation sets varying between three and six samples. PDS with an additive term using five or six samples

produced acceptable prediction errors on each of the three datasets used.

Our model for PDS includes a constant term $\beta_j(1)$. If the spectra in the standardisation sets are mean-centred the constant term will be zero but an adjustment, called an additive background correction (ABC), must be made to the prediction equation to compensate for the mean-centring. Bouveresse et al. (1996) used PDS both with and without the ABC. Their results demonstrate that an ABC is necessary, since in most cases the RMSEP where the ABC is omitted is larger. The model used by Wang and Kowalski (1992) omits the ABC and results in this case are unsatisfactory. They tried to compensate for omission of the ABC by mean-centring the data before starting the process. Surprisingly Wang and Kowalski (1992) reported good results using this method. Possibly in this case the ABC was small, but this will not always be so. Wang et al. (1995) tested a number of methods including using second derivative treatments and mean-centring using an ABC. From their results it was clear that applying pretreatments will not compensate for using an incorrect model: only when an ABC was used were RMSEP acceptable.

A program that performs PDS is available in Wise's PLS Toolbox, (Wise and Gallagher (1998)). We applied PDS to the same two data sets as for SW, NIR and CORN. In each case the RMSEP was calculated for each of the concentrations, before and after standardisation. A window-size of 3 was used, $(k = l = 1)$, with standardisation sets of size 5, 8, and 10. We again give both the RMSEP calculated after standardisation based only on those samples not used in the standardisation process and the RMSEP based on all the samples available. Results are given in tables 3.6 and 3.7. Clearly PDS is better than no standardisation at all and the larger the standardisation set the better the results.

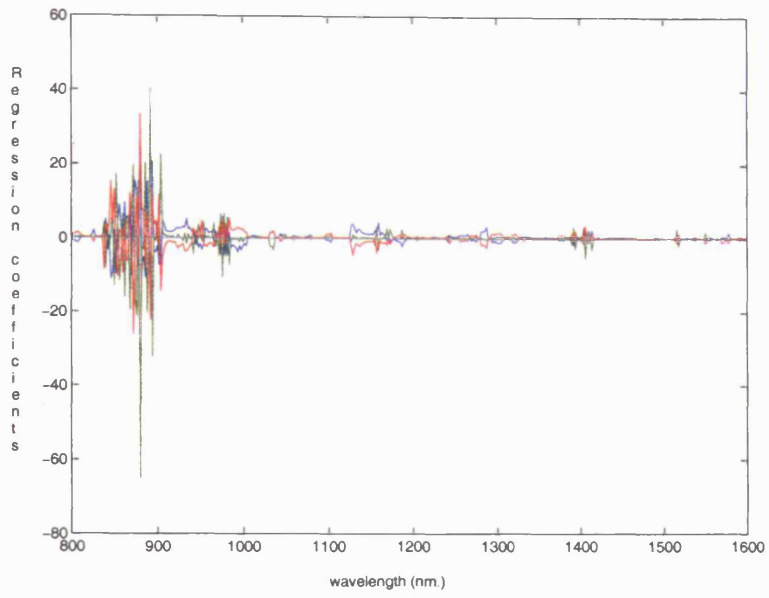
Figures 3.7(a), 3.7(b) and 3.8 show the regression coefficients for the NIR and CORN data. Standardisation sets of size 5 were used for these graphs. Using larger sets gave similar graphs. The graphs suggest that the correlation

Table 3.6: NIR Data before and after applying PDS: RMSEC and RMSEP for dataset omitting standardisation samples and for the entire dataset

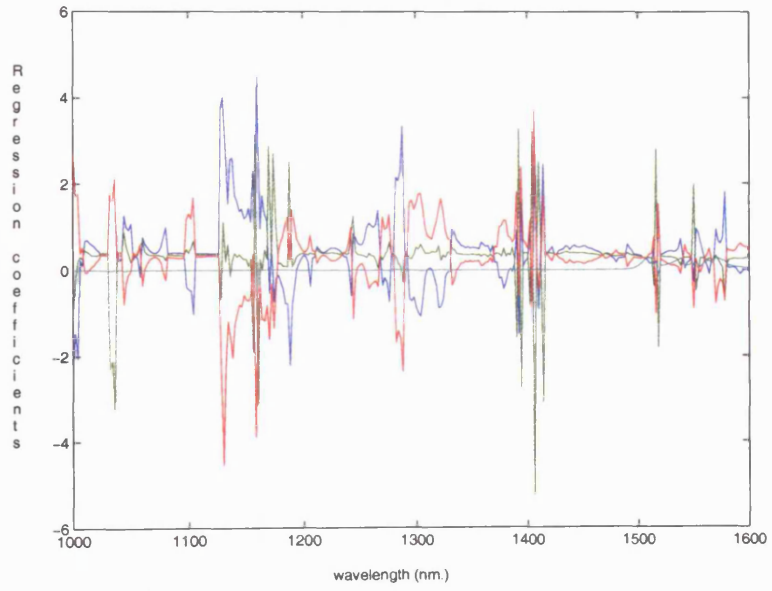
Reference value	1	2	3	4	5
Master	0.6363	0.1185	0.0717	0.1268	0.5696
Slave	9.3995	1.2074	1.4869	2.8324	3.6518
PDS: Standardisation set 5					
25 spectra	1.5734	0.2874	0.2990	0.3616	1.1506
30 spectra	1.4690	0.2895	0.2816	0.3526	1.0749
PDS: Standardisation set 8					
22 spectra	0.9609	0.2737	0.0935	0.1460	0.8272
30 spectra	0.9134	0.2832	0.0914	0.1510	0.7740
PDS: Standardisation set 10					
20 spectra	0.8594	0.2218	0.0984	0.1326	0.7225
30 spectra	0.8148	0.2412	0.0917	0.1376	0.6849

Table 3.7: CORN Data before and after applying PDS: RMSEC and RMSEP for entire dataset and for dataset omitting standardisation samples

Reference value	1	2	3	4
Master	0.0862	0.0986	0.1677	0.2944
Slave	1.4816	0.1672	0.7384	2.1927
PDS: Standardisation set 5				
75 spectra	0.4899	0.1660	0.3122	0.5503
80 spectra	0.4754	0.1640	0.3107	0.5441
PDS: Standardisation set 8				
72 spectra	0.3704	0.1137	0.1716	0.3750
80 spectra	0.3548	0.1113	0.1712	0.3640
PDS: Standardisation set 10				
70 spectra	0.3031	0.1077	0.1677	0.3523
80 spectra	0.3009	0.1066	0.1647	0.3408



(a)



(b)

Figure 3.7: Regression coefficients for NIR data using PDS. (a) for entire wavelength range (800 - 1600 nm. (b) for reduced wavelength range, 1000 - 1600 nm. Intercept-cyan, central coefficient-green, side coefficients-red and blue

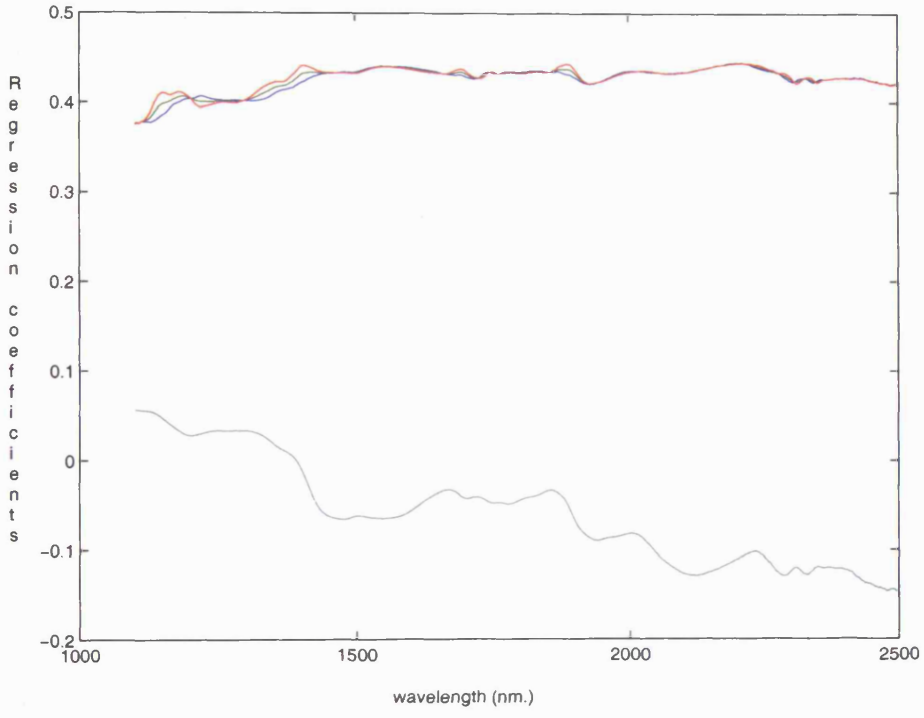


Figure 3.8: Regression coefficients for CORN data using PDS. Intercept-cyan, central coefficient-green, side coefficients-red and blue

between absorbances at adjacent wavelengths has a considerable effect. In the NIR case the first and third coefficients are close to being mirror images of each other. At wavelengths where there is no information in the spectra or the information is confusing the coefficients are unacceptably large. The spectra can be seen more clearly in figure 3.7(b) in which the wavelength range between 800 and 1000 nm where there is little information is omitted. In the CORN graph, each of the three coefficients is about the same size, each contributing approximately one third to the total absorbance. In neither case do the coefficients suggest that the difference between spectra on different instruments consists of a wavelength shift and an absorbance shift, as described in section 3.3.

3.3.3 Methods that adjust the predicted concentration

The slope/bias method (Osborne and Fearn (1983)) is described as adjusting predictions, although once the adjustment has been found it can be combined with the calibration equation so is also seen as a method which adjusts the calibration. The method requires concentrations of interest predicted on both master and slave instruments using the calibration equation from the master instrument. The slave concentrations are regressed on the master concentrations, to give skew and bias constants a and b.

$$Y_m = a + Y_s b$$

Here Y_m and Y_s are the vectors of predicted concentrations on the slave and master instruments. a and b are then used to correct subsequent predictions for samples measured on the slave instrument. The method is very simple and requires only a few samples. However it will work only when slave and master instruments are of the same type and ajustments are small (Fearn (2001)).

3.4 Comparison of standardisation methods

Possibly the only serious comparison of different standardisation methods is due to Wang et al. (1991). The methods compared included PDS, S-W and DS. These are compared with recalibration of the slave instrument using only the standardisation set and methods based on classical and inverse calibration models which almost amount to recalibration and are unlikely to work in practice (Fearn (2001)). The comparison was made using simulated data based on combinations of two pure analytes with spectra on the slave instrument being derived from the master spectra by a non-linear vertical shift and a quadratic horizontal shift, varying between +2 and -2 channels. They used as standardisation sets subsets of the calibration set of sizes between 3 and 10, selected using their method described above. PDS was found to out-perform all other methods for both analytes except where the standardisation set was small, in spite of the fact that it was used without an additive term.

Wang and Kowalski (1992), in a separate study, compared PDS with and without mean-centring. They also considered various pretreatments. Given a small number of samples with known references values, one possibility is to recalibrate the slave instrument using just these samples. Wang and Kowalski (1992) suggest that this procedure is unlikely to give acceptable results.

It is clear from our results for the NIR data in tables 3.4, 3.5 3.6 and 3.7 that PDS works better than SW, though using Gaussian smoothing in the wavelength correction step in SW is a considerable improvement on the original method and leads to results that are better than those found using PDS. The CORN data appears to have at most a small wavelength shift so it is not surprising that results for SW and PDS are very similar.

Chapter 4

Some Bayesian Theory

4.1 Bayesian Paradigm

We discussed in chapter 2 the problems encountered when using multiple linear regression, due to the matrix X^TX being either singular and hence not invertible or being ill-conditioned so that the inverse, when it exists, is unstable. In NIR spectroscopic applications, because X^TX is likely to be singular LSR will not offer a solution. The Bayesian approach to regression avoids the problem of inadequate data by using prior knowledge of the problem. In Bayes theory we begin with a probability distribution on the parameters describing ones' prior state of belief about the data. We then use the data to amend or update the prior distribution to give a posterior distribution which takes into account both our prior knowledge and the data. The Bayesian method has two advantages over the methods described in Chapter 2; it avoids the possibility that no solution can be found and it also enables us to incorporate our prior knowledge into the solution.

Suppose that we wish to describe uncertainty about a parameter, θ, (which may be vector-valued), then a probability distribution for θ, given data X, is given by Bayes' rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{4.1}$$

Here $P(.)$ denotes the probability of the event contained in parentheses, | stands for 'conditional on', X represents the data and θ is a parameter of interest. $P(\theta)$ is a probability distribution representing our prior knowledge about θ, $P(X|\theta)$ is the likelihood of the data, X; that is, the probability of the data conditional on the value of θ. $P(\theta|X)$ is the posterior distribution of the θ; that is, the distribution of θ based on our prior knowledge and the data. The probability of the data, $P(X)$, is often omitted from (1) giving the unnormalised posterior:-

$$P(\theta|X) \propto P(X|\theta)P(\theta) \tag{4.2}$$

The normalisation constant can be calculated though in many situations this is unnecessary.

4.2 Prior distributions

To perform Bayesian inference we need to specify prior distributions for parameters used in the likelihood. The use of prior information is the most controversial aspect of Bayesian analysis, since the prior information is likely to be subjective.

One frequently used method for choosing prior distributions is the selection of priors which when combined with the likelihood of the data yield posterior distributions that belong to the same family as the prior. These are known as conjugate priors. One advantage of this approach is that because the posterior, $P(\theta|X)$, is a known distribution inference can be made directly about the parameter θ. A second advantage is that the information contained in the prior can be identified as additional data. For example, the conjugate prior for the mean of normally distributed data is also normal and the mean of the posterior is a weighted average of the mean of the data and the prior mean. The conjugate prior for the variance of normally distributed data is the inverse-Wishart distribution. The use of the inverse-Wishart as prior illustrates one

of the drawbacks of using conjugate priors. In this case the prior is often insufficiently flexible to represent our prior knowledge completely.

When there is no prior information a diffuse or non-informative prior may be used. Bayes (1763) first used the non-informative prior $P(\theta) = 1$ on the interval [0,1], when estimating θ, the probability of success, in a binomial experiment. The prior expresses the fact that any allowable value of θ is equally likely.

One problem with non-informative priors is that they are often improper in the sense that they are do not have finite integrals and when combined with the likelihood can lead to improper posterior distributions. A uniform prior on an unbounded interval is an example of this. The problem can be avoided by using prior distributions whose variances are large, but finite. An alternative is not to specify the parameters of a prior distribution, but to place priors on them, leading to a hierarchical model.

Bayesian models may be very complex, involving parameters, many of which occur in the joint posterior distribution, but are of no interest in themselves. Their values may be known or may be estimated. These parameters are known as nuisance parameters. An alternative to estimation is to integrate out the nuisance parameter. Suppose μ is a parameter of interest, X the data and σ a nuisance parameter, then we can use

$$P(\mu|X) = \int P(\mu|\sigma, X)P(\sigma|X)d\sigma$$

where integration is over all possible values of σ.

Where nuisance parameters are low-dimesional exact integration or numerical methods may be possible. For high-dimensional situations these methods will be impractical. In this case Markov chain Monte Carlo (MCMC) integration is often used.

4.3 MCMC

There are many variations and an extensive literature on the subject of Markov chain Monte Carlo. Most methods are based on the work of Metropolis et al. (1952) and generalised by Hastings (1970). Essentially their method is to create an approximate distribution for the unknown parameters and alternately sample from and modify this distribution creating a Markov chain whose equilibrium distribution is the joint posterior distribution of the parameters. The sequence of samples generated by this process is defined so that each sample depends only on the previous sample and the sequence of samples obeys the Markov condition and hence forms a Markov chain. The sampling procedure is defined so that the detailed balance equations hold and the stationary distribution is the joint distribution of the unknown parameters. The period until convergence is reached is known as the burn-in. Once the sequence has converged, samples can be used to estimate statistics for the parameters.

The Metropolis-Hastings algorithm is designed to create a sequence of samples, $\theta_1, \theta_2, \ldots$ from the joint distribution of the parameter of interest, $P(\theta|X)$. To do this, a distribution $g(\cdot|\cdot)$ is chosen so that it is possible to sample a candidate, θ^* from $g(\cdot|\theta_t)$. The sample, θ^*, is accepted with probability $\alpha(\theta_t, \theta^*)$ where

$$\alpha(\theta_t, \theta^*) = min(1, \frac{P(\theta^*|X)g(\theta_t|\theta^*)}{P(\theta_t|X)g(\theta^*|\theta_t)})$$

If θ^* is accepted, $\theta_{t+1} = \theta^*$, otherwise, $\theta_{t+1} = \theta_t$. The sampling method guarantees that the sequence $\theta_1, \theta_2, \ldots, \theta_t, \ldots$ forms a Markov chain. The acceptance probability α is constructed so that the detailed balance equations are satisfied and the Markov chain converges to the target distribution, $P(\theta|X)$.

One of the most commonly used special cases of the Metropolis-Hastings algorithm is the Gibbs sampler, so named by Geman and Geman (1984), though the method was in use before this. The Gibbs sampler works by sampling from and updating distributions for separate components of the vector of unknown parameters conditional on the other parameters. Let $(\theta_1, \theta_2, \ldots, \theta_r)$ represent

a set of components of the vector of parameters and denote by $\theta_i(t)$ the value assigned to θ_i at the tth iteration. Let $\theta_{-i}(t) = (\theta_1(t), \theta_2(t), \ldots, \theta_{i-1}(t), \theta_{i+1}(t-1), \ldots, \theta_r(t-1))$ represent the values of the current set of components at time t, omitting θ_i. We start the process by selecting initial values for the θ_i. To determine the value $\theta_i(t)$ we sample from the full conditional distribution, $\pi(\theta_i|\theta_{-i}(t))$. We then use the sampled value to replace the previously used value for θ_i in each of the conditional distributions.

The process of sampling and updating of parameter values is repeated many times, at each stage using the most recently obtained sample values for the parameters in the conditional distributions. To generate samples from the marginal distribution of a parameter of interest, the sequence of samples from the parameter of interest is selected. Sufficient of the initial samples are discarded to allow the chain to have converged. The remaining samples are used to define the distribution of the parameter of interest.

Besag (1974) showed that the joint distribution of the components is determined by the full conditional distributions for the separate components. The same result can be used to establish the detailed balance equations. From this we deduce that the Markov chain derived from the Gibbs sampler is stationary and hence converges to the target distribution.

We need to ensure that sufficient samples have been discarded so that the sequence has converged. The simplest method for assessing convergence is to plot the time series of parameters of interest and check that these appear to be random samples from a fixed distribution. This method can be misleading (Gelman (1996)).

A more objective method, recommended by Gelman and Rubin (1992), is based on analysis of variance. For this method several MCMC sequences for each parameter are generated using widely different starting points. We denote the set of sequences for one such parameter by ψ_{ij} where $i = 1, \ldots, m$ indexes sequences and $j = 1, \ldots, 2n$ indexes samples within a sequence. The

first n samples are discarded and we test whether the remaining sequences have converged to the target distribution. This is done by calculating, two estimates of the variance of the distribution of each parameter. The first, $W = \frac{1}{m(n-1)} \sum_{i,j} (\psi_{ij} - \bar{\psi}_{i.})^2$, will be an under-estimate of the variance of the sequence for finite n and the second, $B = \frac{n}{m-1} \sum_{i=1}^{m} (\bar{\psi}_{i.} - \bar{\psi}_{..})^2$, will overestimate the variance if the sequences have not converged to the target distribution. By considering an estimate of the variance based on a single sequence, it can be shown that $\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{1}{n} B$ is an unbiased estimate of the variance of each of the sequences if they all have converged to the target distribution, otherwise it will over-estimate the variance. The ratio of $\hat{\sigma}^2$ and W is then used to check convergence. If this quantity is close to 1 for each parameter of interest the process can be assumed to have converged.

4.4 Bayesian regression

Let $Y_1, Y_2, \ldots, Y_m$ and $X_1, X_2, \ldots, X_n$ be random variables and assume a regression model

$$(Y_1 Y_2 \ldots Y_m) = (X_1 X_2 \ldots X_n) B + E.$$

Here B is an $n \times m$ matrix of regression coefficients and E is an error matrix. Using the notation of Dawid (1981), $E \sim \mathcal{N}(I, \Sigma)$ with I an identity matrix and Σ an $m \times m$ variance-covariance matrix. In the Bayesian context, priors must be placed on B and Σ. If Σ is assumed known, then the conjugate prior for B is matrix normal. With Σ unknown, the conjugate prior for B will be matrix-normally distributed and dependent on Σ :

$$B - B_0 \sim \mathcal{N}(\Gamma, \Sigma)$$

while Σ will have an inverse-Wishart distribution:

$$\Sigma \sim \mathcal{IW}(\delta, \Sigma_0)$$

The disadvantages of using the conjugate prior here are firstly that the prior variance for B is a scalar multiple of Σ and secondly because the prior expectation for Σ is $\Sigma/(\delta - 2)$ only a single scalar remains to describe the prior variability of Σ, consequently the inverse-Wishart distribution is often insufficiently flexible to describe prior knowledge adequately. An alternative is to assume that B is independent of Σ. We can also assume that Σ is structured in some way - for example that Σ is diagonal or $\Sigma = \tau^2\Sigma_0$ where Σ_0 is constant. We can then place priors on the scalar variables, leading to a non-conjugate analysis.

Brown and Makelainen (1992) explain the principle of structural coherence as applied to selecting priors for a sequenced vector. Since a spectrum of absorbances forms a sequenced vector as defined by Brown and Makelainen (1992), this principle is relevant to our choice of prior.

Several workers, for example George and McCulloch (1993), Brown et al. (1998), structure the prior covariance matrix for regression coefficients as DRD where R has ones on the diagonal and describes the correlation of the regression coefficients while D is a diagonal matrix whose entries give the standard deviations of the coefficients. In our work, we adopt this form of prior.

Chapter 5

Standardisation using Bayesian techniques

5.1 Introduction

For both the Shenk-Westerhaus and Piecewise Direct Standardisation the absorbance correction is performed wavelength by wavelength with no account being taken of the fact that the spectra vary smoothly over the wavelength range and that the regression coefficients are likely to behave in the same way. As mentioned in Chapter 3, the correlation between adjacent wavelengths tends to result in over-fitting, implying a greater wavelength shift than is in fact the case. This can be controlled by shrinking the regression coefficients towards zero intercept and gradient 1. In this chapter we apply Bayesian techniques, using these ideas, to evaluate β.

5.2 Models

The model for both PDS and SW can be written, using the notation of Dawid (1981) as

$$X_m - X_s B - 1_n \alpha^T \sim \mathcal{N}(I_n, \Gamma)$$

Here X_m and X_s are mean-centred $n \times w$ matrices whose rows are spectra for samples in the standardisation set for the master and slave instruments, I_n is an $n \times n$ identity matrix, Γ a $w \times w$ covariance matrix, B is a $w \times w$ matrix of regression coefficients and α is a $w \times 1$ vector of intercepts. For SW, a wavelength adjustment, if needed, is assumed to have been made to X_s, and B is a diagonal matrix. For PDS, B has non-zero elements only on or close to the diagonal.

In a Bayesian approach we need to specify prior distributions for α, B and Γ. To specify the prior for α and B we first reformulate the problem, replacing α and B by a vector, β, containing only α and the entries of B that can be non-zero; $\beta = (\beta_1^T, \beta_2^T, \ldots, \beta_w^T)^T$ where $\beta_j = (\alpha_j, B(j, j-l), \ldots, B(j, j+l))^T$ and replacing X_m by the vector X_M formed by stringing the columns of X_m. i.e.

$$X_M = (X_m(1,1), X_m(2,1), \ldots, X_m(n,1), X_m(1,2), X_m(2,2), \ldots, X_m(n,w))^T$$

X_s is replaced by the block diagonal matrix,

$$X_S = \begin{pmatrix} X_1 & 0 & & \ldots \\ 0 & X_2 & & \ldots \\ \ldots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & X_w \end{pmatrix}$$

where for a window of width $2l+1$,

$$X_r = [1_n X_s(:, r-l) \ldots X_s(:, r+l)]$$

X_S is an $nw \times 2(l+1)w$ matrix. Where $j = r - l \leq 0$ or $j = r + l > w$ so that $X_s(:, j)$ was undefined, adjacent values, $X_s(:, 1)$ or $X_s(:, w)$ were used. The first and last wavelengths were not used in the calibration for PDS. For SW, $l = 0$, while for PDS, we set $l = 1$. Windows of width greater than 1 are possible, as are asymmetric windows, but for our data, which has wavelength intervals of 2 nanometers, we found that $l = 2$ did not improve predictions.

The model for the data is :-

$$X_M \sim N(X_S\beta, \Sigma)$$

where

$$\Sigma = \Gamma \otimes I_n.$$

Here the variance-covariance matrix, Γ, is assumed diagonal, Γ =diag$(\sigma_1^2, \sigma_2^2, \ldots, \sigma_w^2)$. This is clearly an over-simplification, but is useful in simplifying the inversion of the posterior variance.

The likelihood is given by

$$(2\pi)^{-nw/2}|\Sigma|^{-1/2}\exp\{-\frac{1}{2}(X_M - X_S\beta)^T\Sigma^{-1}(X_M - X_S\beta)\} \qquad (5.1)$$

5.3 Prior distributions

Initially we estimated Σ, the variance of the model, from the data. The conjugate prior for β for the normal likelihood with known variance is also normal. With

$$\beta \sim N(\beta_0, \Sigma_0)$$

the prior probability density function for β is

$$(2\pi)^{-w(2l+2)/2}|\Sigma_0|^{-1/2}\exp\{-\frac{1}{2}(\beta - \beta_0)^T\Sigma_0^{-1}(\beta - \beta_0)\}. \qquad (5.2)$$

We select the value for β_0, the prior mean for β, to be the value β would take in the model if X_M were identical to X_S.

For SW we have

$$\beta_0 = (0, 1,\ 0, 1, \ldots, 0, 1)^T,$$

and for PDS with a window of width 3,

$$\beta_0 = (0, 0, 1, 0,\ 0, 0, 1, 0, \ldots, 0, 0, 1, 0)^T.$$

Σ_0 is taken to be of the form:-

$$\Sigma_0 = \tau^2\Sigma_\rho \otimes \Sigma_1 \qquad (5.3)$$

with Σ_1 a $(2l+2) \times (2l+2)$ diagonal matrix, Σ_0 is a $2(l+1)w \times 2(l+1)w$ matrix. The correlation matrix, Σ_ρ, is based on a first order autoregressive structure AR(1), i.e.

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{w-1} \\ \rho & 1 & \rho & \dots & \rho^{w-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho^{w-1} & \rho^{w-2} & \dots & \rho & 1 \end{pmatrix}.$$

Σ_ρ is a correlation matrix, with ρ, $(\rho < 1)$, specifying the prior correlation between absorbances at adjacent wavelengths. With ρ close to 1 these are, in the prior, highly correlated, while as the distance, $2d$ nm, between wavelengths increases the correlation, given by ρ^d, diminishes. $\tau^2\Sigma_1$ specifies the prior variances of the coefficients of β. There is some justification for using different values for the diagonal elements of Σ_1, allowing for different variances for the intercept and gradient coefficients of β. One might also allow different values for the correlation coefficients, ρ, corresponding to different components of β, leading to a different form for Σ_0. We investigated these alternatives without finding any clear advantages to compensate for the added complication, so used the same value of ρ and $\Sigma_1 = I_2$ or I_4 leading to the form given in (5.3). The effect of using this prior is to shrink coefficients towards β_0 and, with ρ close to 1, to smooth β so that coefficients relating to adjacent wavelengths will be similar in value.

5.4 Posterior distribution

The posterior probability density function for β is, using Bayesian theory, the product of the prior for β given by equation 5.2 in section 5.3 and the likelihood, given by equation 5.1 in section 5.2.

$$\beta | X_M, X_S, \Sigma, \Sigma_0, \beta_0 \sim N(B_1^{-1}b, B_1^{-1}) \tag{5.4}$$

where $B_1 = X_S^T \Sigma^{-1} X_S + \Sigma_0^{-1}$ and $b = X_S^T \Sigma^{-1} X_M + \Sigma_0^{-1} \beta_0$. This result was part of a more general result proved by Lindley and Smith (1971). The mean, $B_1^{-1} b$, is a weighted average of the weighted least squares regression coefficient and β_0, the prior mean for β.

The main problem with using this result for standardisation is the need to invert the matrix, B_1, a $(2l+2)w \times (2l+2)w$ matrix with $w = 401$ and $w = 700$ for our two datasets. Matrices of this size are far too large to invert using brute force.

Because Σ and Σ_1 are diagonal their inverses are easily calculated. $\Sigma_0^{-1} = \tau^{-2} \Sigma_\rho^{-1} \otimes \Sigma_1^{-1}$ and

$$\Sigma_\rho^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix}$$

To invert B_1 we exploit the fact that the matrix involved is highly structured and almost block diagonal.

We use the identity

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1} \end{pmatrix}$$

where $E = (D - CA^{-1}B)$, to invert the matrix a block at a time using an inductive process. B_1 is a $(2l+2)w \times (2l+2)w$ matrix of the form

$$\begin{pmatrix} Z_1 & P & 0 & 0 & 0 & \dots \\ P & Z_2 & P & 0 & 0 & \dots \\ 0 & P & Z_3 & P & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \dots & Z_w \end{pmatrix}$$

with $Z_i = X_i^T X_i / \sigma_i^2 + \tau^{-2}((1+\rho^2)/(1-\rho^2))\Sigma_1^{-1}$, for $2 \le i \le w-1$, $Z_i = X_i^T X_i / \sigma_i^2 + (\tau^{-2}/(1-\rho^2))\Sigma_1^{-1}$, for $i = 1$ and w, and $P = -\rho\tau^{-2}\Sigma_1^{-1}/(1-\rho^2)$,

$(2l+2) \times (2l+2)$ matrices. We first invert Z_1, using a Matlab subroutine then set $A = Z_1$ and use this and the above identity to invert

$$\begin{pmatrix} Z_1 & P \\ P & Z_2 \end{pmatrix}.$$

Once this matrix has been inverted we set

$$A = \begin{pmatrix} Z_1 & P \\ P & Z_2 \end{pmatrix}$$

and use the result to invert

$$\begin{pmatrix} Z_1 & P & 0 \\ P & Z_2 & P \\ 0 & P & Z_3 \end{pmatrix},$$

again using the identity, and so on. Using this method only $(2l+2) \times (2l+2)$ matrices need to be inverted directly.

5.5 Choice of plug-in parameter values

τ, $\sigma_1, \sigma_2, \ldots, \sigma_w$ and ρ are unknown parameters . ρ is close to but less than 1. We estimated ρ by minimising $(X_M - X_S\hat{\beta})^T(X_M - X_S\hat{\beta})$ for the standardisation set. Values of 0.98 and 0.99 appeared to work well for the NIR data and smaller values - 0.9 or 0.91- for the CORN data. We can estimate $\sigma_1, \sigma_2, \ldots, \sigma_w$ from the data, replacing β by the regression coefficient, b, found in SW or PDS:-

$$s_j^2 \propto (X_m(:,j) - X_j b_j)^T (X_m(:,j) - X_j b_j). \tag{5.5}$$

Using regression coefficients found in SW or PDS will lead to an under-estimate of the s_j^2 but this is unimportant since, because of the form of the posterior mean, which is used as our estimate of β, we need only the ratio of the σ_i to τ so do not need the constant of proportionality in expression 5.5. The usefulness

of (5.5) is that it gives an estimate of the relative sizes of the s_j^2. Selecting a value for τ is more difficult. τ^{-2} is the weight given to β_0 in the expression for β, so τ^{-2} must be large enough to ensure sufficient shrinkage towards β_0, without allowing β to be dominated by β_0. One method for selecting τ is to use the value that minimises the RMSEP for the standardisation set. The problem with this approach is that, because the standardisation set is selected to include as much variability as possible, its members are in some sense the most extreme samples, so that the τ that works well for them will not necessarily be the best value for more typical samples. This proved to be the case for our data. An alternative is to base the value of τ on samples that are close to the mean of the spectra and so more representative of those to be standardised. We used the three samples closest to the mean of the master set and selected the value for τ that minimised the RMSEP on the corresponding three samples on the standardised slave spectra. Using this value of τ reduced the RMSEP on the slave spectra after standardisation, but at the expense of using three extra samples for standardisation.

So that comparisons could be made with other methods results reported, here we use parameters selected by minimising the RMSEP on the standardisation set only.

5.6 MCMC

An alternative to using plug-in values for the variances, τ^2 and σ_j^2, is to place priors on them and use Gibbs sampling to do a full Bayes analysis.

5.6.1 Priors

The conjugate prior distribution for the precision of a normal distribution is the gamma distribution. τ^{-2} and the σ_j^{-2} were given independent gamma

priors:-

$$\tau^{-2} \sim ga(\alpha, \gamma).$$

$$\sigma_j^{-2} \sim ga(a, as^2)$$

The prior distribution for σ_j^{-2} has mean s^{-2} and standard deviation $s^{-2}/\sqrt{a}$. With parameters in this form we can select s^2 using our knowledge of the values of the plug-in parameters s_j^2 and make a small to give a diffuse prior.

5.6.2 Full conditional distributions

To use Gibbs sampling we need to sample from the full conditional distributions of each of the parameters. The full conditional for β is given by equation 5.4 in section 5.4. The full conditional distributions for τ^{-2}, $\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_w^{-2}$ are all gamma distributions depending on the data and their prior distributions.

$$\tau^{-2}|\alpha, \gamma, data, \beta \sim ga[\alpha + (l+1)w, \gamma + \frac{1}{2}(\beta - \beta_0)^T \Sigma_\rho^{-1} \otimes \Sigma_1^{-1}(\beta - \beta_0)],$$

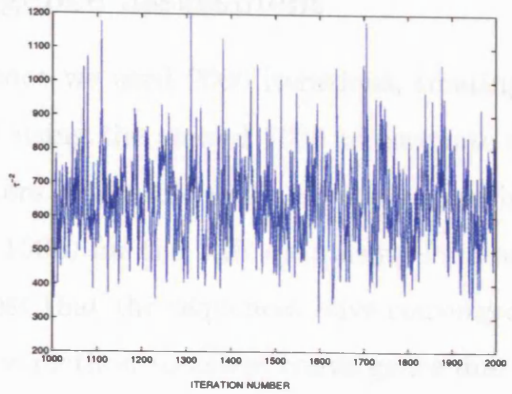
$$\sigma_j^{-2}|a, s, data, \beta_j \sim ga[a + \frac{n}{2}, as^2 + \frac{1}{2}\mu(:,j)^T\mu(:,j)]$$

where $\mu(:,j) = X_m(:,j) - X_j\beta_j$ and β_j is the component of β at wavelength j.

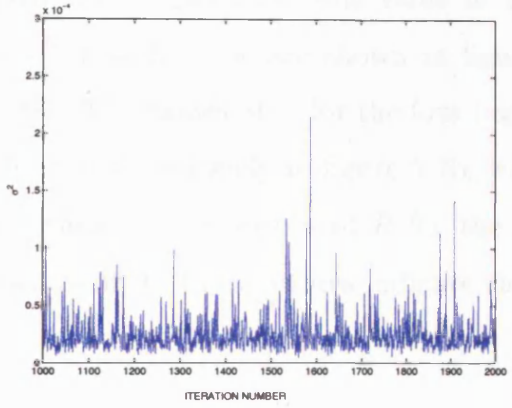
5.6.3 Parameter specification

There remain four parameters that must be specified: α, γ, a and s.

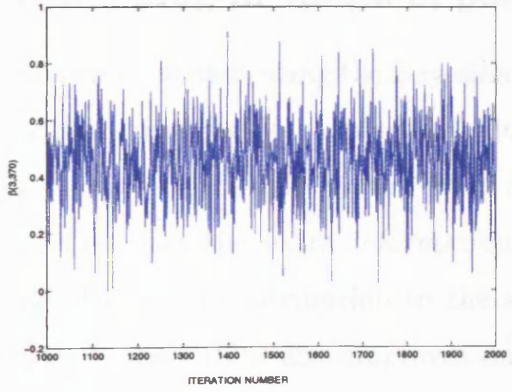
Since there is little information available from the data to determine τ, its value is highly dependent upon the parameters for its prior, in particular on the value of α. If α is too small, τ^{-2} converges to zero resulting in no shrinkage or smoothing of β; if α is too large, β will be dominated by its prior, the data not being allowed sufficient weight. We selected a value for α which was as small as possible without allowing τ^{-2} to become zero or lead to noisy regression coefficients. γ is related to the variance of the precision, τ^{-2}; by choosing a small value for γ we ensure that in the prior, τ^{-2} has a large variance. We gave s^{-2} the value 10^{-5}, based on the values of the s_j^{-2} used in the previous section. We set $a = 10^{-2}$ leading to a fairly diffuse prior.



(a)



(b)



(c)

Figure 5.1: Graphs for MCMC sequences a) τ^{-2}, b) σ^2_{1536} and c) $\beta(3, 1536)$

5.6.4 Convergence assessment

To ensure convergence we used 2000 iterations, treating the first 1000 as the burn-in period and using the second 1000 to estimate statistics of the distributions of parameters. Example plots for the second halves of sequences for τ^{-2}, σ^2_{1536} and $\beta(3, 1536)$ for the NIR data using PDS are shown in figure 5.1. These graphs suggest that the sequences have converged.

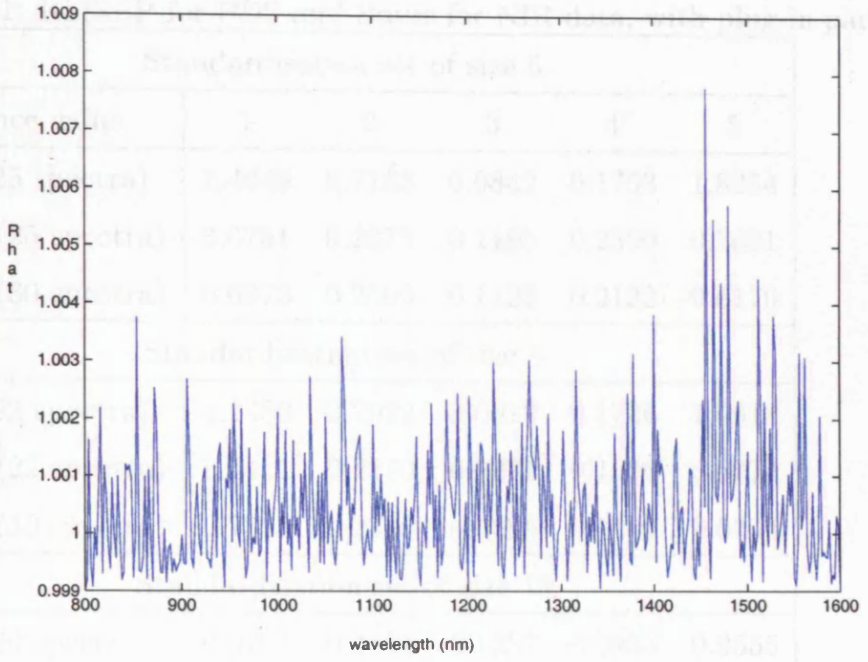
A more objective method to assess convergence due to Gelman and Rubin (1992) and described in section 4.3 was also used. We generated four chains independently using different starting points, and for each parameter, calculated $\hat{R}$. For the NIR data using PDS, the value of $\hat{R}$ for τ^{-2} was 1.0072. The values of $\hat{R}$ for $\sigma^2_j, j = 1, \ldots, w$ are shown in figure 5.2a. They ranged between .999 and 1.007. The values of $\hat{R}$ for the four regression coefficients for the same data are shown consecutively in figure 5.2b, with $\hat{R}$ for the intercept first, showing largest values. The values of $\hat{R}$ for the regression coefficients range between .999 and 1.004. These values indicate that the sequences have converged.

5.7 Results and discussion

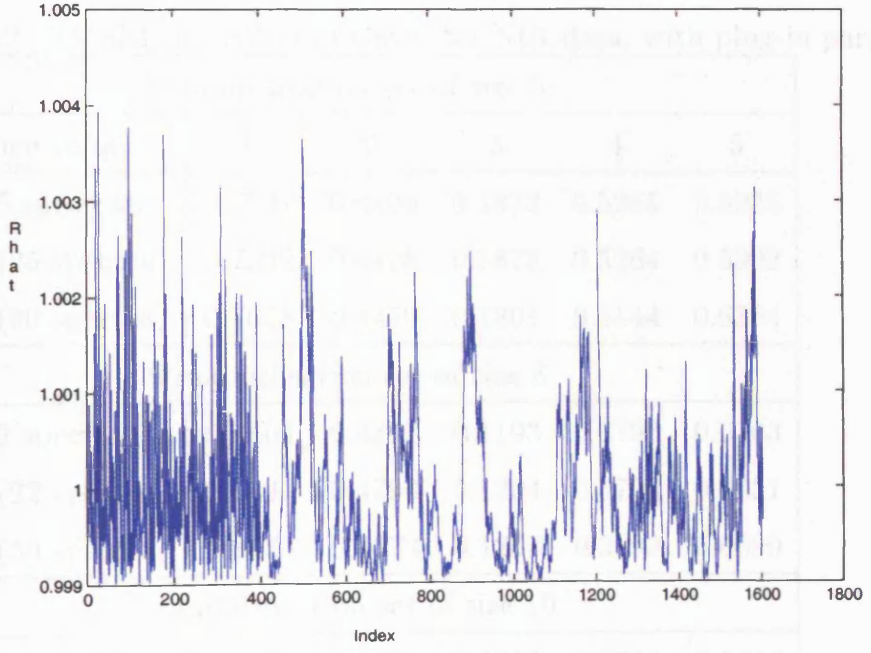
5.7.1 Results using plug-in values of parameters

From tables 5.1 and 5.3 it is clear that using the Bayesian method with the PDS model, even with plug-in parameters, is a substantial improvement over PDS alone. The greatest reduction in RMSEP occurs when only 5 standardisation samples are used. In this case the extra information in the form of prior distributions makes a substantial contribution to the solution. Using larger standardisation sets Bayes with PDS also improves on using PDS alone for both datasets but the improvement is less pronounced.

For the NIR data, Gaussian smoothing is applied to the correlation coefficients before the wavelength shift is estimated both for the SW method and



(a)



(b)

Figure 5.2: Values of $\hat{R}$ for a) σ and b) β, with values for $\beta(1,:)$ shown between 1 and 400, $\beta(2,:)$ between 401 and 800 etc.

Table 5.1: RMSEP for PDS and Bayes for NIR data, with plug-in parameters

Standardisation set of size 5					
Reference value	1	2	3	4	5
PDS (25 spectra)	2.4049	0.7163	0.0842	0.1753	1.8254
Bayes (25 spectra)	0.6751	0.2673	0.1195	0.2300	0.5691
Bayes (30 spectra)	0.6973	0.2509	0.1122	0.2132	0.6170
Standardisation set of size 8					
PDS (22 spectra)	1.2483	0.2922	0.0919	0.1726	1.0618
Bayes (22 spectra)	0.7411	0.2170	0.0871	0.1326	0.6014
Bayes (30 spectra)	0.7537	0.2320	0.0896	0.1480	0.6204
Standardisation set of size 10					
PDS (20 spectra)	0.9701	0.4671	0.1237	0.3983	0.9555
Bayes (20 spectra)	0.6826	0.1773	0.0982	0.1349	0.5519
Bayes (30 spectra)	0.6976	0.2014	0.0925	0.1385	0.5793

Table 5.2: RMSEP for SW and Bayes for NIR data, with plug-in parameters

Standardisation set of size 5					
Reference value	1	2	3	4	5
SW (25 spectra)	0.7537	0.4406	0.1873	0.5255	0.5925
Bayes (25 spectra)	0.7532	0.4428	0.1878	0.5264	0.5902
Bayes (30 spectra)	0.7658	0.4479	0.1801	0.5144	0.6364
Standardisation set of size 8					
SW (22 spectra)	0.7464	0.4850	0.1193	0.3798	0.6103
Bayes (22 spectra)	0.7441	0.4796	0.1204	0.3799	0.6051
Bayes (30 spectra)	0.8033	0.4673	0.1207	0.3833	0.6590
Standardisation set of size 10					
SW (20 spectra)	0.6571	0.2420	0.1729	0.2534	0.5604
Bayes (20 spectra)	0.6586	0.2386	0.1728	0.2524	0.5587
Bayes (30 spectra)	0.7513	0.2707	0.1719	0.2613	0.6191

(Gaussian smoothing is applied to the correlation coefficients found using SW)

Table 5.3: RMSEP for PDS and Bayes for CORN data, plug-in parameters

Standardisation set of size 5				
Reference value	1	2	3	4
PDS (75 spectra)	0.4907	0.1646	0.3167	0.5112
Bayes (75 spectra)	0.3148	0.1022	0.1613	0.3609
Bayes (80 spectra)	0.3089	0.1036	0.1621	0.3622
Standardisation set of size 8				
PDS (72 spectra)	0.4006	0.1110	0.1915	0.4219
Bayes (72 spectra)	0.3014	0.1046	0.1505	0.3223
Bayes (80 spectra)	0.2919	0.1041	0.1496	0.3171
Standardisation set of size 10				
PDS(70 spectra)	0.3293	0.1057	0.1881	0.3898
Bayes (70 spectra)	0.2707	0.1030	0.1533	0.3144
Bayes (80 spectra)	0.2718	0.1025	0.1501	0.3136

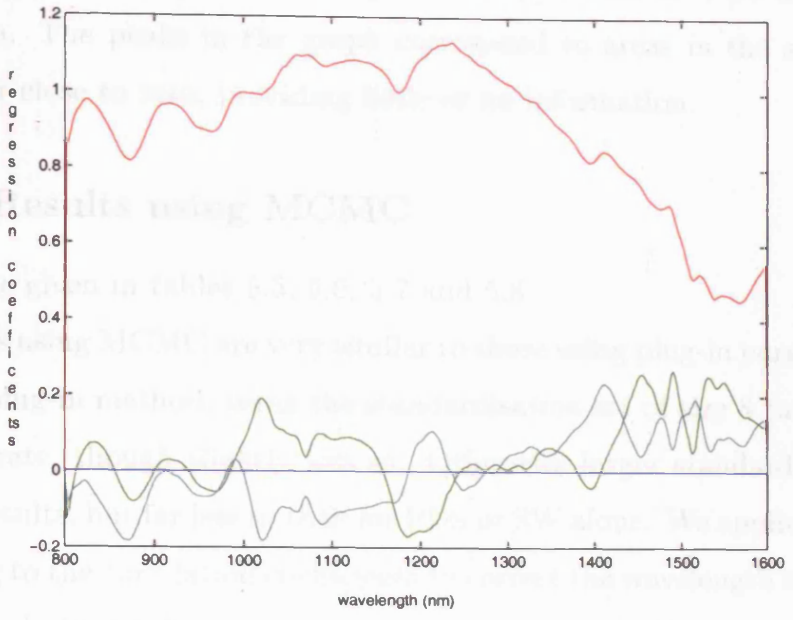
Table 5.4: RMSEP for SW and Bayes for CORN data, with plug-in parameters

Standardisation set of size 5				
Reference value	1	2	3	4
SW (75 spectra)	0.4947	0.1685	0.3067	0.5321
Bayes (75 spectra)	0.3714	0.0990	0.1668	0.3830
Bayes (80 spectra)	0.3606	0.0970	0.1653	0.3842
Standardisation set of size 8				
SW (72 spectra)	0.3761	0.1132	0.1713	0.3768
Bayes (72 spectra)	0.3677	0.0990	0.1588	0.3891
Bayes (80 spectra)	0.3506	0.0965	0.1543	0.3746
Standardisation set of size 10				
SW (70 spectra)	0.3070	0.1075	0.1672	0.3552
Bayes (70 spectra)	0.3003	0.0979	0.1568	0.3543
Bayes (80 spectra)	0.2968	0.0960	0.1504	0.3483

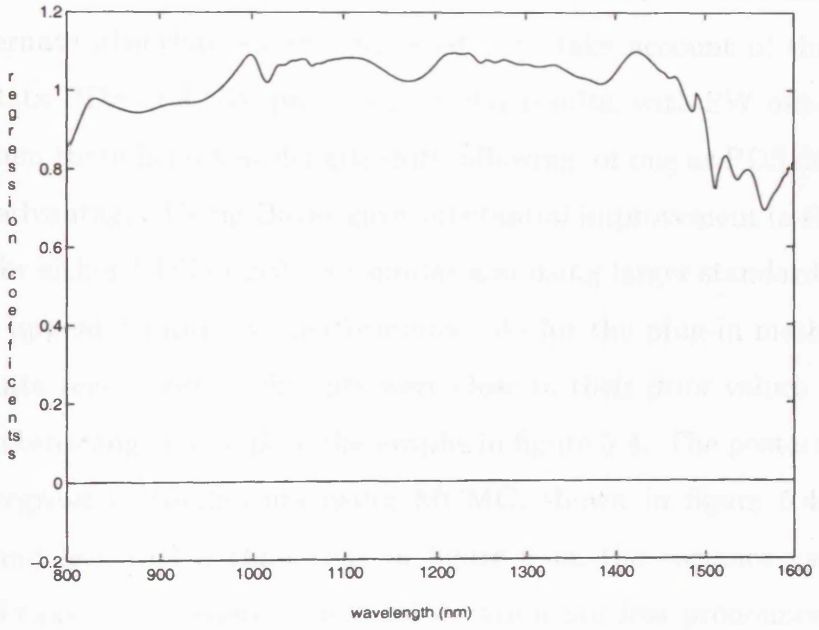
for the Bayesian method using the SW model. It is evident from table 5.2 that the Bayesian method is no better than the wavelength by wavelength regression used in SW in the case where an accurate wavelength shift has already been applied. This is slightly surprising since comparing regression coefficients for SW (figure 3.6(a)) with those for Bayes using the SW model in figure 5.3(b) it is clear that the Bayes solution is better, though the main improvement is in the area at the lower end of the spectrum where the information is sparse. Although SW with Gaussian smoothing of the correlation coefficients produced smaller RMSEP than PDS, when Bayes was used with each of these two models, PDS outperformed SW. The advantage of SW is that, unlike PDS, it corrects the actual errors - the wavelength and absorbance shifts - directly. The information added in the Bayes solution more than compensates for this when using PDS, but is less important for SW.

For the CORN data PDS and SW produce very similar results (tables 5.3 and 5.4). This is probably due to the fact that there is at most a very small wavelength shift. However when Bayes is used PDS again out-performed SW.

The effect of shrinking the regression coefficients can be seen if the graphs in figure 5.3 are compared with those in figures 3.7(a) and 3.6(a). The wild variation occurring when PDS or SW alone is used is controlled by the use of Bayesian methods. More significantly, in the PDS graphs, it can be seen that the effects of correlations between absorbances at adjacent wavelengths are reduced by using Bayesian shrinkage and smoothing so that more weight is given to absorbances on the slave instrument at the same wavelength as that being modelled on the master instrument (corresponding to the regression coefficients shown in red in figure 5.3 (a)) and less to those at adjacent wavelengths (represented by the regression coefficients shown in green and cyan). The effects of correlations are still evident in the symmetry exhibited between the coefficients adjacent to the central one. Note that because the data are mean-centred, the intercept is always close to zero. Graphs of the variance of



(a)



(b)

Figure 5.3: Graphs of regression coefficients for NIR data for a) PDS (with central coefficient-red and side coefficients-green and cyan) and b) SW, using Bayesian regression with plug-in parameter values, on mean-centred data

the posterior distribution for the regression coefficients for PDS are shown in figure 5.4a. The peaks in the graph correspond to areas in the spectra that are zero or close to zero, providing little or no information.

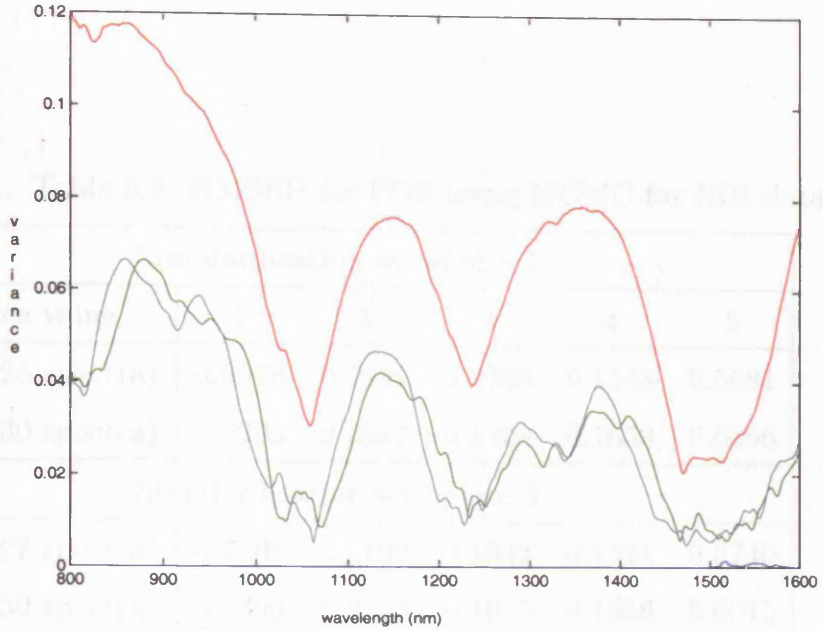
5.7.2 Results using MCMC

Results are given in tables 5.5, 5.6, 5.7 and 5.8.

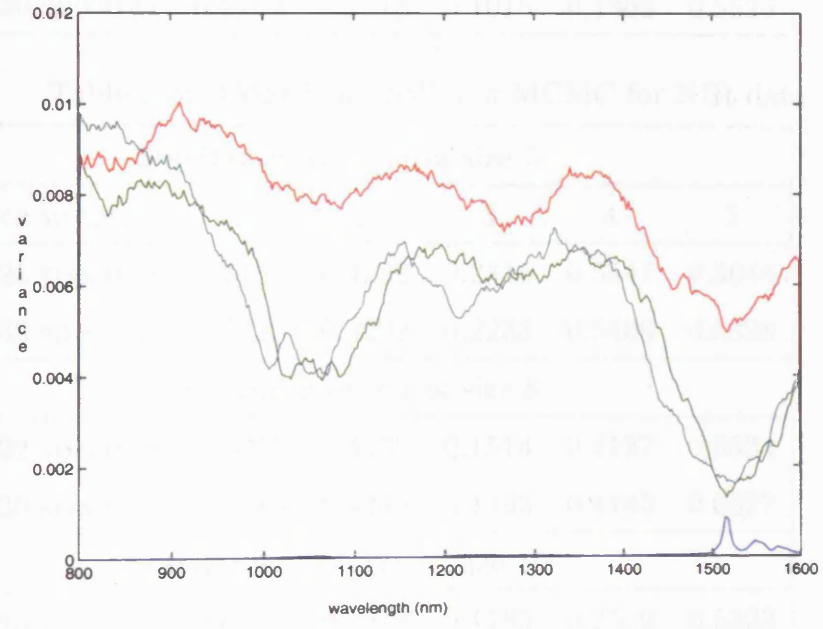
Results using MCMC are very similar to those using plug-in parameters. As with the plug-in method, using the standardisation set of size 8 causes results to deteriorate, though slightly less so. Otherwise larger standardisation sets improve results, but far less so than for PDS or SW alone. We applied Gaussian smoothing to the correlation coefficients to correct the wavelength shift for SW and although this worked much better than using quadratic smoothing, results for the NIR data were still inferior to those obtained using PDS.

Because of the time required to run MCMC on the CORN data, we used only alternate absorbances and adjusted ρ to take account of this. For the CORN data PDS and SW produced similar results, with SW out-performing PDS. When there is no wavelength shift, allowing for one as PDS does is possibly a disadvantage. Using Bayes gave substantial improvement in the RMSEP. Results for either PDS or SW are similar and using larger standardisation sets does not appear to improve performance. As for the plug-in method, for the CORN data regression coefficients were close to their prior values.

It is interesting to compare the graphs in figure 5.4. The posterior variance for the regression coefficients using MCMC, shown in figure 5.4b, is much smaller and less varied than that in figure 5.4a, the variance using plug-in values. Peaks where there is less information are less pronounced, but the minima between 1000 and 1100 nm. and at approximately 1500 nm where the spectra are most informative are well-defined.



(a)



(b)

Figure 5.4: Variance of the posterior distribution for the regression coefficients for the NIR data using PDS. a) using plug-in values, b) using MCMC. Variance for the intercept is shown in blue, the central coefficient in red and the side coefficients in green and cyan

Table 5.5: RMSEP for PDS using MCMC for NIR data

Standardisation set of size 5					
Reference value	1	2	3	4	5
Bayes (25 spectra)	0.6978	0.2636	0.1260	0.1548	0.5681
Bayes (30 spectra)	0.7205	0.2511	0.1269	0.1659	0.6256
Standardisation set of size 8					
Bayes (22 spectra)	0.7208	0.2109	0.1042	0.1521	0.5740
Bayes (30 spectra)	0.7393	0.2313	0.1045	0.1656	0.6015
Standardisation set of size 10					
Bayes (20 spectra)	0.6617	0.1760	0.1047	0.1413	0.5252
Bayes (30 spectra)	0.6764	0.2018	0.1015	0.1508	0.5525

Table 5.6: RMSEP for SW and MCMC for NIR data

Standardisation set of size 5					
Reference value	1	2	3	4	5
Bayes (25 spectra)	0.7451	0.4222	0.2315	0.5541	0.5648
Bayes (30 spectra)	0.7823	0.4292	0.2233	0.5409	0.6328
Standardisation set of size 8					
Bayes (22 spectra)	0.7057	0.4370	0.1514	0.4137	0.5634
Bayes (30 spectra)	0.7816	0.4283	0.1493	0.4145	0.6327
Standardisation set of size 10					
Bayes (20 spectra)	0.6501	0.2274	0.1785	0.2569	0.5323
Bayes (30 spectra)	0.7560	0.2629	0.1779	0.2661	0.6084

Table 5.7: RMSEP for PDS using MCMC for CORN data

Standardisation set of size 5				
Reference value	1	2	3	4
Bayes (75 spectra)	0.2820	0.0998	0.1494	0.3218
Bayes (80 spectra)	0.2796	0.1022	0.1497	0.3244
Standardisation set of size 8				
Bayes (72 spectra)	0.2805	0.0997	0.1501	0.3590
Bayes (80 spectra)	0.2783	0.1022	0.1506	0.3589
Standardisation set of size 10				
Bayes (70 spectra)	0.2840	0.1001	0.1478	0.3240
Bayes (80 spectra)	0.2813	0.1023	0.1482	0.3252

Table 5.8: RMSEP for SW and MCMC for CORN data

Standardisation set of size 5				
Reference value	1	2	3	4
Bayes (75 spectra)	0.2588	0.0993	0.1446	0.3062
Bayes (80 spectra)	0.2570	0.1018	0.1453	0.3135
Standardisation set of size 8				
Bayes (72 spectra)	0.2773	0.1024	0.1492	0.3058
Bayes (80 spectra)	0.2690	0.1029	0.1493	0.3073
Standardisation set of size 10				
Bayes (70spectra)	0.2551	0.1019	0.1496	0.3055
Bayes (80 spectra)	0.2562	0.1022	0.1483	0.3059

Chapter 6

Dynamic linear modelling

6.1 Introduction

Strictly speaking our data do not form a time series, but as they are sequential they can, by replacing time with wavelength, be treated as one, enabling us to use the extensive theory of time-series analysis.

Time series arise in practice in many fields. In economics they are used to trace the variation in prices, incomes and other economic predictors. In meteorology they are used to record and analyse variables such as rainfall and temperature used in forecasting. They are used to record and predict demographic changes and in many other fields where sequential data exist.

One very successful approach to time series was developed in the engineering literature by Kalman (1963). Kalman proposed a dynamic linear model (DLM) - linear in that at any time, t, the model for observed data y_t is a linear regression model based on known covariates, F_t, given by

$$Y_t = F_t^T \theta_t + v_t \tag{6.1}$$

and dynamic in that the regression coefficient, θ_t, is updated via the recursion

$$\theta_t = G_t \theta_{t-1} + w_t \tag{6.2}$$

Here v_t and w_t are random variables for which only wide sense assumptions

are made; i.e. assumptions about means and variances, but no distributional assumptions. In these equations the observed data, Y_t, is modelled as being linearly related to known covariates, F_t, via regression coefficients, θ_t, with error, v_t, while the regression coefficients at time t are related to those at time $t-1$ through matrix G_t, assumed known, again with an error term, w_t.

Kalman's proofs are given in terms of projections of the response variable onto the subspace spanned by the columns of the explanatory matrix. From his dynamic linear model he derived recursive updating equations for the time-dependent regression coefficients and proved that the estimated regression coefficients were unbiased and minimum mean square linear estimates (MMSLE). The recursive equations give an estimate of the mean and variance of θ_t, the regression coefficient at time t, based on all information available up to and including time t, in terms of means and variances of random variables calculated at the previous stage and observations up to time t. He also formulated equations that predict the values of future regression coefficients in terms of observations and estimated variables. Finally, by means of a second recursive relation, Kalman estimated retrospectively the expected value and variance of each of the estimated regression coefficients conditional on all the available data both past and future. These last estimates are made using the "Kalman smoother" equations.

Because the Kalman filter and smoother equations were developed in the engineering literature and in terms not necessarily familiar to statisticians they were not widely used by statisticians.

Duncan and Horn (1972) explained and proved Kalman's results in terms of regression theory, showing that the estimated regression coefficients are both MMSLE and minimum variance unbiased linear estimators and, in a Bayesian context, with the stronger assumptions that errors are normal and uncorrelated, that the estimated regression coefficients, $\theta_t|Y_t$ are Bayes posterior means or minimum mean square error estimates.

Harvey and Phillips (1979) showed that the autoregressive-moving average process can be expressed as a dynamic linear model. Consequently the recursive equations developed by Kalman can be used to establish distributional results and in forecasting and retrospective analysis.

Harrison and Stevens (1976) developed dynamic linear modelling in a Bayesian framework. They showed that essentially all ARMA-type models of classical linear time-series models can be treated as special cases of the dynamic linear model. In West and Harrison (1997) Kalman smoother equations are developed. West and Harrison (1997) emphasise the advantages of the use of dynamic linear modelling compared with previous time-series methods, in particular that seasonality and non-stationarity can be dealt with within the dynamic linear framework, the possibility of intervention by adjusting the size of the error w_t if conditions change and probably most important the recursive nature of the process makes it computationally efficient.

DLM theory can be combined with generalised linear models (GLMs) developed by Nelder and Wedderburn (1972) allowing the observations, Y_t, to follow non-normal distributions from the exponential family of distributions with natural parameter non-linearly related to regression coefficients via a monotonic link function. By using conjugate priors, closed form versions of the updating equations are found. West et al. (1985) give details of this application of DLMs to the theory of GLMs.

6.2 The dynamic linear model

Throughout this section we use the results and notation of West and Harrison (1997) where proofs of the results can be found.

The dynamic linear model is defined by the observation equation,

$$Y_t = F_t^T \theta_t + v_t, \quad v_t \sim N(0, V_t) \tag{6.3}$$

the system equation,

$$\theta_t = G_t\theta_{t-1} + w_t, \quad w_t \sim N(0, W_t) \tag{6.4}$$

and the quadruple

$$\{F_t, G_t, V_t, W_t\}$$

which is assumed known. Here Y_t is the $n \times 1$ vector of observed data at time t, θ_t, the regression coefficient at time t, F_t is an $r \times n$ matrix of covariates, G_t, an $r \times r$ matrix and V_t and W_t $n \times n$ and $r \times r$ covariance matrices. Each of the v_t and w_t are assumed independent of all the others.

Writing D_t for all information available up to and including time t, it can be shown that if

$$(a) \qquad \theta_{t-1}|D_{t-1} \sim N(m_{t-1}, C_{t-1})$$

then

$$(b) \qquad \theta_t|D_{t-1} \sim N(a_t, R_t)$$

where $a_t = G_t m_{t-1}$ and $R_t = G_t C_{t-1} G_t^T + W_t$

$$(c) \qquad Y_t|D_{t-1} \sim N(f_t, Q_t)$$

where $f_t = F_t^T a_t$ and $Q_t = F_t^T R_t F_t + V_t$

$$(d) \qquad \theta_t|D_t \sim N(m_t, C_t)$$

where $m_t = a_t + A_t(Y_t - f_t)$, $A_t = R_t F_t Q_t^{-1}$ and $C_t = R_t - R_t F_t Q_t^{-1} F_t^T R_t$. This result is proved in West and Harrison (1997).

To start the process we need values for m_0 and C_0. The distributions of the state, θ_t, and the observation, Y_t, $t = 1, \ldots$ can then be calculated recursively. This process is known as the Kalman filter.

One of the main uses of time series is to forecast the distribution of future values $Y_{t+k}|D_t$, of the time series, $Y_1, Y_2, \ldots, Y_t$. These can be derived from the predicted state distributions,

$$\theta_{t+k}|D_t \sim N(a_t(k), R_t(k))$$

which can be calculated recursively using $a_t(k) = G_{t+k}a_t(k-1)$ and $R_t(k) = G_{t+k}R_t(k-1)G_{t+k}^T + W_{t+k}$ and starting the recursion with $a_t(0) = m_t$ and $R_t(0) = C_t$. We also have

$$Y_{t+k}|D_t \sim N(f_t(k), Q_t(k))$$

where $f_t(k) = F_{t+k}^T a_t(k)$ and $Q_t(k) = F_{t+k}^T R_t(k) F_{t+k}^T + V_{t+k}$.

It is also possible to derive distributions for $\theta_{t-k}|D_t$ for $0 < k < t$. In this case future as well as past information is used to derive the marginal distribution of θ_{t-k}. Using the same notation as for previous results,

$$\theta_{t-k}|D_t \sim N(a_t(-k), R_t(-k))$$

where $a_t(-k) = m_{t-k} + B_{t-k}(a_t(-k+1) - a_{t-k+1})$, $B_t = C_t G_{t+1}^T R_{t+1}^{-1}$ and $R_t(-k) = C_{t-k} + B_{t-k}(R_t(-k+1) - R_{t-k+1})B_{t-k}^T$. The recursion is again started using $a_t(0) = m_t$ and $R_t(0) = C_t$.

This process is known as smoothing the series. In order to apply this result it is necessary to make a first pass through the data applying the Kalman filter, then working backwards through the data, starting with $a_t(0) = m_t$ and $R_t(0) = C_t$ and at each stage using the quantities evaluated at the previous stage and those found in the first pass, the distribution for the smoothed regression coefficients $\theta_{t-k}|D_t$ can be estimated and from them smoothed observational responses.

West and Harrison (1997) also describe a generalisation in the one-dimensional case of the process to the situation where the observational variance is unknown. When the precision is given a gamma prior its posterior is shown also to have a gamma distribution, while the marginal distributions for $\theta_t|D_t$ and $Y_t|D_t$ are Student-t distributions with means having the same form as in the case of known observational variance.

Chapter 7

Application of dynamic linear modelling to standardisation

7.1 Introduction

As mentioned in the previous chapter our data do not form a time series, but because they are sequential, they can be treated as one. The difference between a spectrum and a time series is that while a time series is always ordered in the direction of increasing time, there is no compelling reason to order a spectrum starting at the lower wavelength end. In fact, radiation may also be characterised in terms of the wavenumber, the reciprocal of the wavelength, and in this case the graph of a spectrum may be presented in the reverse order.

PDS and SW fit linear models compatible with the observation equation in DLM and the system equation enables us, by suitable choice of values for W_t to ensure that the regression coefficients vary smoothly from one wavelength to the next.

Using the notation of the previous chapter, for PDS or SW, the observation equation is

$$X_m(:,t) = X_t\theta_t + v_t \quad v_t \sim N(0, V_t)$$

where

$$X_t = [1_n X_s(:, t-l) \ldots X_s(:, t+l)].$$

For SW, $l = 0$ and a wavelength adjustment is assumed to have been applied to X_s. For PDS l is small and for our data $l = 1$ was used. This is essentially the same model as was used for the Bayesian techniques in chapter 5. The observational variance, V_t, is estimated by

$$V_t = k(X_m(:, t) - X_t b_t)^T (X_m(:, t) - X_t b_t)$$

where b_t is the regression coefficient found from PDS or SW. The constant, k combines the divisor for the degrees of freedom in the expression for V_t and a further factor allowing for the fact that V_t is calculated using b_t from PDS or SW for which we expect a smaller variance than for θ_t found using DLM. We use this expression for V_t because it provides a good estimate for the relative sizes of the variances at different wavelengths. We need only specify V_t up to a constant of proportionality since it is only the ratio of V_t to W that occurs in the calculation of θ_t.

The system equation for both PDS and SW is

$$\theta_t = \theta_{t-1} + w_t, \quad w_t \sim N(0, W)$$

We make the assumption that the w_t follow the same normal distribution $N(0, W)$ for all values of t, also that given θ_{t-1}, the components of θ_t are uncorrelated so that W is assumed diagonal.

Using this setup, the observation equation describes $X_m(:, t)$ as being given by $X_t\theta_t$ with v_t allowing for a normally distributed random error in the data. The system equation allows θ_t to vary smoothly from one wavelength to the next, controlled by w_t, a random normally distributed error term.

One problem with this model is that it is first order autoregressive and not stationary leading to steadily increasing marginal variances for $X_M = (X_m(:, 1)^T \; X_m(:, 2)^T \ldots X_m(:, w)^T)^T$. This can be avoided if we replace the system

equation by

$$\theta_t = \lambda\theta_{t-1} + (1-\lambda)\phi + w_t, \quad 0 \leq \lambda < 1$$

where $\phi = (0,0,1,0)^T$ for PDS with window-size 3 and $\phi = (0,1)^T$ for SW. This is again an autoregressive model, but is stationary with $Var(\theta_t) \to W/(1-\lambda^2)$ as long as $\lambda < 1$. Using this system equation also has the advantage that the regression coefficients, θ_t, are shrunk towards the prior value ϕ. In terms of the notation of the previous chapter we define

$$\tilde{\theta}_t = \begin{pmatrix} \theta_t \\ \phi \end{pmatrix} \text{ and } G_t = \begin{pmatrix} \lambda & 1-\lambda \\ 0 & 1 \end{pmatrix} \otimes I_r$$

and our system equation becomes

$$\tilde{\theta}_t = G_t\tilde{\theta}_{t-1} + w_t$$

($r = 2$ for SW and $r = 4$ for PDS).

As previously mentioned, a NIR spectrum is not strictly a time series. Although sequential, there is no reason to start the sequence at the lower frequency end of the spectrum rather than the upper end. If we begin the Kalman filter at the lower frequency end, the resulting regression coefficients, $\theta_t|D_t, t = 1,\ldots,w$, will be different from those obtained by starting at the upper end since the data, D_t, will be different. However the Kalman smoother output, $\theta_t|D_w$ relies on the same data, D_w, irrespective of whether we start at the upper or lower end of the spectrum. Clearly the Kalman smoother output is preferable in our situation since estimates of θ_t are then based on all the data rather than only part of it.

7.2 Choice of parameter values

There are two parameters for which values must be chosen before we can apply DLM to our standardisation problem. These are λ and W. λ is a scalar and if we assume that $W = \mu I_{2l+2}$ we need only to select the value of

the scalar μ. One method for estimating parameters is to select those which lead to minimum values of the RMSEP, $\sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^p)^2}{n}}$, for the standardisation set, where n is the number of samples in the standardisation set, y_i, $i = 1, \ldots, n$ are the reference values for the n samples and y_i^p are the reference values estimated using DLM. Here we encounter the same problem as we did in chapter 5 in selecting prior parameter values, that, because the members of the standardisation set are in some sense the most extreme members of the calibration set they are atypical. An alternative is to use spectra that are closer to the mean to determine the parameters. This produced better results but involves using a larger standardisation set. We also need to select values for m_0 and C_0, the mean and variance of the distribution of θ_0, to start the recursive process. We set $m_0 = \phi$ and $C_0 = W/(1 - \lambda^2)$. We found that using these prior values, the values of $\theta_t|D_w$, $t = 1, \ldots, w$, were the same irrespective of whether the process was started at the lower or upper frequency end of the spectrum.

For the results given in the next section parameters were selected based on minimising the mean RMSEP over the given reference values for the standardisation set used for DLM. So that proper comparisons can be made between results for different sized standardisation sets, we quote the RMSEP for the entire calibration set as well as for the set excluding the standardisation samples.

For the NIR data, results were highly dependent on the values of the parameters. Values in the region of 0.96 to 0.98 for λ appear to give the best results, giving most weight to the data, with only slight shrinkage towards ϕ. For the CORN data parameter values were less critical. Values of λ that were greater than 0.9 gave best results. Values of μ varied from values between 10^{-4} and 10^{-3} for PDS using the NIR data, depending on the size of the standardisation set to much smaller values around 10^{-5} for the CORN data.

Table 7.1: RMSEP for PDS and DLM (Kalman smoother) for NIR data

Standardisation set of size 5					
Reference value	1	2	3	4	5
PDS(25 spectra)	2.4049	0.7163	0.0842	0.1753	1.8254
DLM(25 spectra)	0.7572	0.2471	0.1184	0.2274	0.6807
DLM (30 spectra)	0.7417	0.2388	0.1012	0.2126	0.6897
Standardisation set of size 8					
PDS(22 spectra)	1.2483	0.2922	0.0919	0.1726	1.0618
DLM (22 spectra)	0.7136	0.2011	0.0995	0.1504	0.5802
DLM (30 spectra)	0.6873	0.2237	0.0887	0.1657	0.5495
Standardisation set of size 10					
PDS (20 spectra)	0.9701	0.4671	0.1237	0.3983	0.9555
DLM (20 spectra)	0.6752	0.1722	0.0934	0.1284	0.5480
DLM(30 spectra)	0.6325	0.2026	0.0787	0.1417	0.5066

7.3 Results and discussion

In table 7.1 we give the RMSEP for each response variable for DLM using the PDS model. SW did not work well with the NIR data and these results have not been included. This is probably due to the failure of the wavelength correction step rather than poor perfomance of DLM, since for the CORN data where there is at most a very small wavelength shift, SW with DLM works well even when only five standardisation samples were used. In tables 7.2 and 7.3 we give the RMSEP for each response variable using DLM with the PDS and SW models. Standardisation sets of sizes 5, 8 and 10 samples were used. The comparable results for PDS and SW are also included. From these results it can be seen that DLM is particularly effective when the standardisation set is small, making substantial improvements on PDS or SW. Increasing the sample size improves the performance of DLM for the NIR data but the improvement

over PDS or SW is less pronounced. Surprisingly, increasing the size of the standardisation set from 5 to 8 using the CORN data leads to slightly higher RMSEPs for SW.

The graphs in figure 7.1 are of regression coefficients for PDS using DLM for each dataset. For these 10 samples were used for the standardisation set. Coefficients for smaller standardisation sets were similar. Both graphs suggest that the correlation between absorbances at consecutive wavelengths appears to have affected the coefficients, though to a lesser extent than using PDS alone. In the NIR graph regression coefficients at the two adjacent wavelengths (shown in green and cyan) are often close to being mirror images of each other. In the CORN graph they are almost indistinguishable. These effects can be attributed to the correlations since one would expect only one of the adjacent regression coefficients to be non-zero. Comparing these graphs with figures 3.7 and 3.8, it is clear that the shrinking and smoothing have a considerable effect.

In figure 7.2 we have graphs of the variances of the posterior distributions of the regression coefficients for the NIR data for both the Kalman filter and the Kalman smoother. (The two graphs that are close to zero are the variances for the intercepts for the Kalman filter and smoother). Using the Kalman smoother leads to more accurate results and also results with much smaller variances. For the CORN data, the variance is always very small - less than 10^{-5} for the whole wavelength.

7.4 Remark

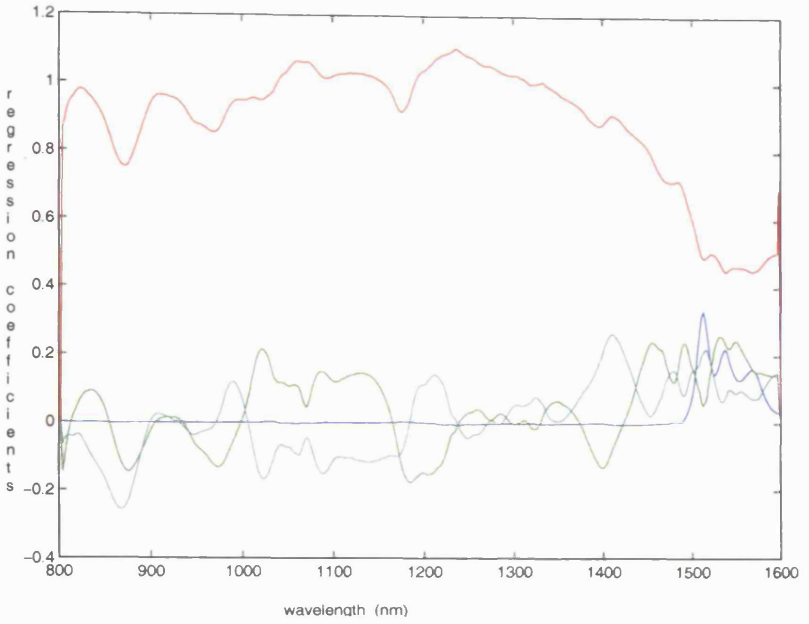
In chapter 11 we make a detailed comparison of all methods examined in this thesis. Here we mention that while RMSEP for DLM are generally only slightly larger than those for the Bayesian method described in chapter 5, DLM is much faster and easier to apply, so should not be dismissed as being inferior to the full Bayesian method.

Table 7.2: RMSEP for PDS and DLM (Kalman smoother) for CORN data

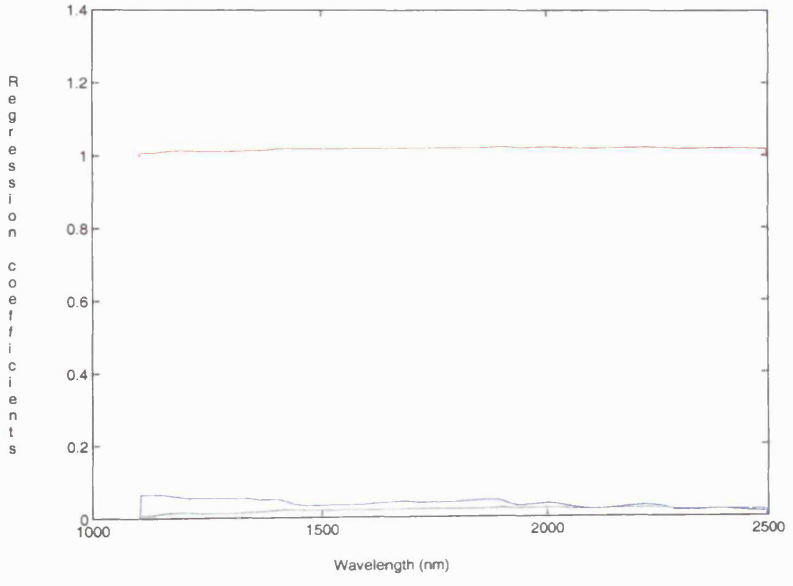
Standardisation set of size 5				
Reference value	1	2	3	4
PDS (75 spectra)	0.4907	0.1646	0.3167	0.5112
DLM (75 spectra)	0.2784	0.1013	0.1590	0.3503
DLM (80 spectra)	0.2762	0.1033	0.1614	0.3446
Standardisation set of size 8				
PDS (72 spectra)	0.4006	0.1110	0.1915	0.4219
DLM (72 spectra)	0.3475	0.1081	0.1643	0.3693
DLM 80 spectra	0.3334	0.1066	0.1637	0.3572
Standardisation set of size 10				
PDS(70 spectra)	0.3293	0.1057	0.1881	0.3898
DLM (70 spectra)	0.2724	0.1030	0.1541	0.3459
DLM (80 spectra)	0.2734	0.1027	0.1512	0.3345

Table 7.3: RMSEP for SW and DLM (Kalman smoother) for CORN data

Standardisation set of size 5				
Reference value	1	2	3	4
SW (75 spectra)	0.4947	0.1685	0.3067	0.5321
DLM (75 spectra)	0.3941	0.1070	0.2604	0.4526
DLM (80 spectra)	0.3862	0.1077	0.2626	0.4519
Standardisation set of size 8				
SW (72 spectra)	0.3761	0.1132	0.1713	0.3768
DLM (72 spectra)	0.3830	0.1061	0.1855	0.3659
DLM (80 spectra)	0.3677	0.1043	0.1840	0.3558
Standardisation set of size 10				
SW(70 spectra)	0.3070	0.1075	0.1672	0.3552
DLM (70spectra)	0.3266	0.1046	0.1791	0.3510
DLM (80 spectra)	0.3250	0.1039	0.1759	0.3414



(a)



(b)

Figure 7.1: Graphs showing regression coefficients for PDS using DLM a)NIR data, b) CORN data. Intercept-blue, central coefficient-red, side coefficients-green, cyan

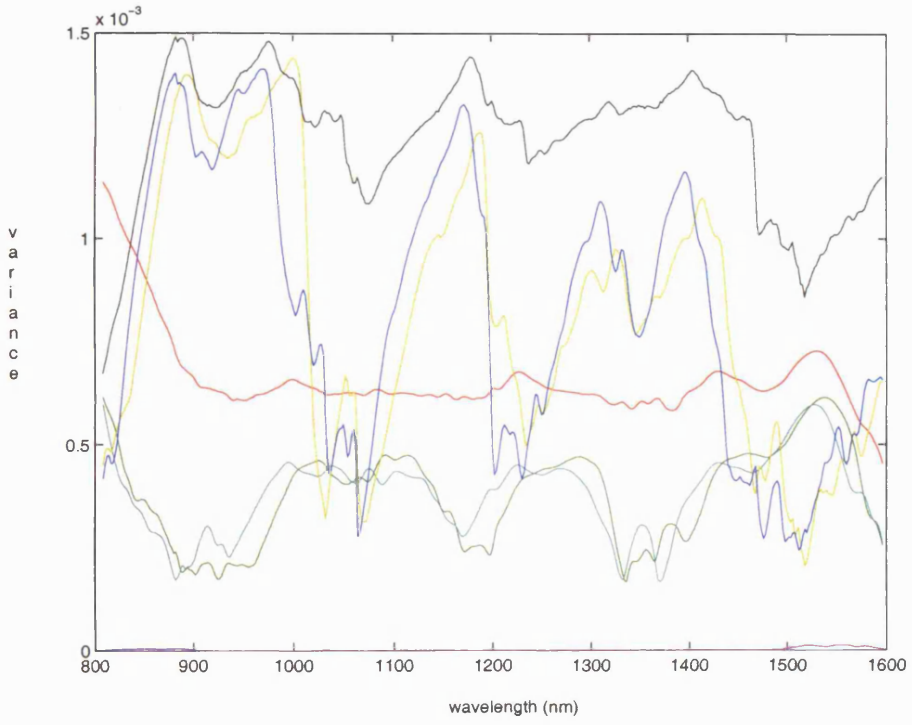


Figure 7.2: Variance of regression coefficients using the Kalman filter, intercept-magenta, central coefficient-grey, side coefficients-yellow and blue and Kalman smoother, intercept-blue, central coefficient-red, side coefficients-green and cyan: NIR data

Chapter 8

Fourier transforms and wavelets

8.1 Fourier Transforms

In 1807 the French mathematician J-B Fourier announced his discovery that any function $f(x) \in L^2(\mathcal{R})$ defined on (or periodic on) the interval $[-\pi, \pi]$ can be represented as a linear combination of the functions $\sin(nx)$ and $\cos(nx)$, $n = 0, 1, \ldots$

i.e.

$$f(x) = a_0 + \sum_{n=1}^{n=\infty} a_n \cos(nx) + \sum_{n=1}^{n=\infty} b_n \sin(nx)$$

where the coefficients a_n, b_n are given by $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)dx$ and $b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx$. The functions $\frac{1}{\sqrt{\pi}}\cos(nx)$, $\frac{1}{\sqrt{\pi}}\sin(nx), n = 0, 1, \ldots$ form an orthonormal set on $[-\pi, \pi]$ i.e.

$$\int_{-\pi}^{\pi} \sin(mx)\cos(nx)dx = 0,$$

$$\int_{-\pi}^{\pi} \sin(mx)\sin(nx)dx = 0, \int_{-\pi}^{\pi} \cos(mx)\cos(nx)dx = 0, \ n \neq m$$

and

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \sin(mx)^2 dx = \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(nx)^2 dx = 1.$$

The idea can be adapted to apply to functions defined on intervals other than $[-\pi, \pi]$. It can also be used where the function f is defined on a finite number of evenly spaced points. Integrals are then replaced by finite sums. In

many cases the coefficients a_n and b_n for large n are small and can be omitted without affecting greatly the representation of the function f. This allows a parsimonious representation of the function in terms of its Fourier coefficients. The development of the Fast Fourier Transform by Cooley and Tukey (1965), which provides a rapid method for calculating Fourier coefficients, has resulted in increasing use of Fourier transforms for describing and summarising signals.

8.2 Wavelets

8.2.1 Introduction

One of the problems with Fourier representations of functions is that they do not deal well with local features. Wavelets are families of orthonormal functions which because of the local nature of the basis functions can define different parts of the signal at different levels of resolution, and hence represent local features parsimoniously. A family of wavelet functions is defined in terms of a multiresolution analysis, a sequence of function spaces, each one contained within the succeeding one and each allowing a more detailed description than the previous one. From these is defined an orthonormal basis which spans various function spaces. The basis functions are all translations and dilations of the same compactly supported functions. To represent the signal in the wavelet domain it is first decomposed into approximation and detail at the lowest level, level 1. The detail is specific to a small neighbourhood while the approximation is a slightly smoothed version of the original signal. The approximation and detail each require half as many coefficients as the original signal. The approximation can then be decomposed to give a higher level approximation and a coarser level of detail. This process can be continued until the approximation is represented by a single scaling function.

Although the first wavelet basis was discovered by Alfred Haar in 1910, the use of wavelets has developed rapidly only since 1988 with the publication

by Daubechies of a paper describing a family of wavelets which are compactly supported and have varying degrees of smoothness (Daubechies (1988)).The wavelet coefficients for a function are convolutions of the function with the basis functions, so that an integration must be performed to estimate each coefficient. Mallat (1989) devised a method of calculating wavelet coefficients using recurrence relations linking coefficients at one level to those at the one above. Mallet's cascade method greatly reduces the time required to transform a function to the wavelet domain. Daubechies' wavelets and Mallat's cascade algorithm have enabled wavelets to be used extensively, in particular, for signal denoising, image analysis and in data compression.

8.2.2 Discrete Wavelet Bases

As with Fourier transforms, when functions are defined only at discrete points their wavelet representation takes the form of a finite linear combination of basis elements. For a formal definition of wavelets defined in terms of a multiresolution analysis, see, for example Ogden (1997) or Vidakovic (1999). For our purposes a brief summary is all that is required. A multiresolution analysis is a sequence of self-similar subspaces of $L_2(\mathcal{R})$,

$$\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \ldots$$

for which there exists a scaling function, ϕ, which with its translates, $\phi(. - k)$, $k \in \mathcal{Z}$, forms an orthonormal basis for V_0. From the scaling function, we derive a function ψ, the mother wavelet, which together with its translations and dilations

$$\psi_{lk}(x) = 2^{l/2}\psi(2^l x - k), l, k \in \mathcal{Z}$$

forms an orthonormal basis for $L_2(\mathcal{R})$. Here l represents the level of detail and k the offset. For a fixed value of l, ψ_{lk} forms an orthonormal basis for the difference space

$$W_l = V_{l+1} \setminus V_l. \tag{8.1}$$

By an appropriate choice of mother wavelet, ψ, a wide variety of functions can be represented in the wavelet domain. From 8.2.2 we deduce that

$$V_0 = V_{-N} \oplus \sum_{-N}^{-1} W_i.$$

If we have a function, f, defined only at integer points $j = 1, 2, \ldots, 2^N$ it can be represented in the wavelet domain by coefficients u and w_{lk}, $l = -1, -2, \ldots, -N$, $k = 0, 1, \ldots, 2^{N+l} - 1$:-

$$f(x) = u\phi(x) + \sum_{l=-N}^{-1} \sum_{k=0}^{2^{N+l}-1} w_{lk} 2^{l/2} \psi_{lk}(x).$$

Since ϕ, the basis element for V_{-N}, and the ψ_{lk} form an orthonormal basis for $L_2(\mathcal{R})$ there exists an orthogonal matrix W which transforms the vector $\underline{f} = (f(1)\ f(2) \ldots f(2^N))$ to $(u, w_{-N\,0},\ w_{-N+1\,0}, \ldots, w_{-1\,2^{N-1}-1})$.

8.3 Thresholding

The key to data compression and signal denoising in the wavelet domain is thresholding. Suppose $\underline{g} = (g(1), g(2), \ldots, g(n))^T$ is the vector of values for an observed function $g(.)$, which is the result of the signal $\underline{f}$ and added white noise. Then our model is

$$\underline{g} = \underline{f} + \underline{\epsilon}$$

where $\underline{\epsilon} \sim N(\underline{0}, \sigma^2 I_n)$. Because of the orthogonality of W, in the wavelet domain each wavelet coefficient of the observed signal, $\hat{w}_{lk}$, is the sum of the wavelet coefficient of the true signal, w_{lk}, and an error term, ϵ_{lk}, where $\epsilon_{lk} \sim N(0, \sigma^2)$ and independent. Thresholding is an attempt to determine and remove ϵ_{lk} and hence recover the true signal.

Donoho and Johnstone (1994) proposed two types of wavelet thresholding, hard and soft. In hard thresholding, all wavelet coefficients whose absolute value is less than a fixed threshold, λ, are set to zero.

$$\delta^h(x, \lambda) = \begin{cases} 0 & if \quad |x| \leq \lambda \\ x & if \quad |x| > \lambda \end{cases}$$

In soft thresholding all small coefficients are again set to zero, but the remaining coefficients are reduced in absolute value by λ.

$$\delta^s(x,\lambda) = \begin{cases} 0 & if \quad |x| \leq \lambda \\ sign(x)(|x| - \lambda) & if \quad |x| > \lambda \end{cases}$$

The threshold may vary with the resolution level, l. Donoho and Johnstone (1994) showed that the universal threshold $\sqrt{2\log n}\ \sigma$ has good asymptotic properties. In using this threshold, σ, which is usually unknown must be estimated from the w_{lk}. Data are compressed by converting them to the wavelet domain and then setting some coefficients to zero. In the wavelet domain small coefficients may be assumed to be due to noise and ignored. The result is a function from which noise has been eliminated and which is also efficiently represented.

8.3.1 Bayesian thresholding

Bayesian methods may be used to shrink estimates towards a suitably chosen prior mean so are a natural choice for implementing thresholding rules. In the wavelet domain our model is:-

$$\hat{w}_{lk} \sim N(w_{lk}, \sigma_l^2)$$

We require priors for w_{lk} and σ_l^2. Prior knowledge can be used in specifying these and by choosing the appropriate form for the prior for w_{lk} the posterior model will correspond to either a shrinkage or a thresholding rule.

Hard thresholding is equivalent to the variable selection problem in linear regression, i.e. selecting variables from $X_1, \ldots, X_p$ when $Y|\beta, \sigma^2 \sim N([X_1 \ldots X_p]\beta, \sigma^2)$, $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$. George and McCulloch (1993) tackled this problem by placing a prior on β_j given by

$$\beta_j|\gamma_j, c_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2), \quad \gamma_j \sim Ber(p_j)$$

with the τ_j small and c_j large ($c_j > 1$). The effect of this prior is to make β_j small when $\gamma_j = 0$ so that the corresponding variable can be ignored, while

if $\gamma_j = 1$, β_j may be large and the corresponding variable will be included in the model. Variable selection was then made by generating sequences of the unknown parameters using Gibbs sampling and observing the most frequently selected variables. Clyde et al. (1998) used this same prior for w_{lk} when estimating wavelet coefficients. Estimation of the w_{lk} using the posterior mean in this case leads to a shrinkage rule for the w_{lk}, not a thresholding rule.

In a paper in which a Bayesian approach is compared to other thresholding methods Abramovich et al. (1998) used as prior for w_{lk}

$$w_{lk} \sim \pi_l N(0, \tau_l^2) + (1 - \pi_l)\delta(0)$$

where $\delta(0)$ is a point mass at 0. They prove that using this prior and estimating w_{lk} by the posterior median leads to a true thresholding rule. In their paper they assumed τ_l^2 to be of the form $2^{-al}C_1$ and π_l of the form $\min(1, 2^{-bl}C_2)$. With parameters of this form, the number of non-zero wavelet coefficients is related to the value of b, with $b = 0$ giving every wavelet coefficient an equal probability of being zero, while $b = 1$ implies that the expected number of non-zero coefficients is the same at each level. Abramovitch et al. show that with this form of parametrisation of the prior for w_{lk}, the values of a and b determine the particular Besov space to which the modelled function belongs. Thus a and b determine the regularity of the function.

8.4 Image analysis

Image analysis is a form of signal denoising where the signal is a two-dimensional image, for example a digital picture to which noise has been added. The situation is usually defined in terms of arrays of pixels, S, and a 'colouring' of S denoted by $x = (x_1, x_2, \ldots, x_n)$ where x_i is the colour of the ith pixel. A probability distribution $\{p(x)\}$ which assigns colourings to S is defined. If $p(x_i)$, the distribution on x_i, is conditional only on the values of $x_{\delta i}$ where δ_i

is the set representing a neighbourhood of i i.e.

$$p(x_i|x_{S-\{i\}}) = p(x_i|x_{\delta i})$$

then $\{p(x)\}$ in known as a locally dependent random field.

The problem of de-noising an image has been tackled using a Bayesian approach by Geman and Geman (1984) and by Besag (1986). The true image, x, is described in terms of the observed image, y, by

$$p(x|y) \propto l(y|x)p(x).$$

Here $l(y|x)$ is the likelihood of the observed image, y, given the true image, x. A further simplifying assumption, that given the colouring of the true image, the observed images y_i are independent, gives $l(y|x) = \prod_i l(y_i|x_i)$

The number of pixels is usually very large and the computational problem correspondingly large. Geman and Geman (1984) used simulated annealing and Gibbs sampling to find the posterior mode for $p(x|y)$. For each i, x_i is sampled from the conditional density

$$P_T(x_i|y, \hat{x}_{S-\{i\}}) \propto [l(y_i|x_i)p_i(x_i|\hat{x}_{\delta i})]^{1/T}$$

where $\hat{x}$ represents the estimated value of x, updated after each sampling, and T is initially large and changes according to a defined schedule which eventually converges to a value close to zero.

Besag (1986) proposed a simpler procedure, iterated conditional modes (ICM), where, for each i in turn, the value of x_i which maximises

$$p(x_i|y, x_{S-\{i\}}) = l(y_i|x_i)p_i(x_i|\hat{x}_{\delta(i)})$$

is substituted in $\hat{x}_S$ and the procedure continued until convergence is reached.

8.5 Deformable templates

A different problem in image analysis was introduced by Grenander (1970). In this the true image or template, F, and the observed image, G are assumed

known and the emphasis is on finding the deformation function that transforms F to G. This results in points in the template being identified with the corresponding points in the image. Amit (1994) defined the deformation function, ϕ, with $\phi(x) = x + U(x)$ as the function that minimises

$$J(U) \propto E(U,U) + \frac{1}{2}\int |F(x+U(x)) - G(x)|^2 dx$$

Here E is a bilinear form that penalises non-smooth functions. U is parametrised in terms of 2-dimensional Fourier or wavelet coefficients. Fourier and wavelet bases have considerable computational advantages here since the problem can first be solved using lower level coefficients only, and then the resolution increased. Since the Fourier and wavelet bases are orthogonal, the penalty term, E, takes a particularly simple form. The wavelet solution has the advantage over using Fourier transforms that local changes can be made without affecting other regions.

In a Bayesian framework, a prior is placed on the deformation function. Aykroyd and Mardia (1996) tackled a 1-dimensional deformation problem using the Metropolis algorithm to estimate the deformation function which was modelled in the wavelet domain with hard thresholding imposed on the coefficients.

Downie et al. (1996) used a similar approach when analysing deformations of the femoral condyle (part of the human thigh bone). They used as a prior on the wavelet coefficients,

$$w_s \sim (1-\pi_l)N(0,\sigma_l^2) + \pi_l\delta(0)$$

where w_s is a 2-dimensional wavelet coefficient at resolution level l. This again leads a thresholding rule.

When investigating the same problem Downie (1997) used three different methods to estimate wavelet coefficients. These included steepest ascent, which in their application was equivalent to ICM, and simulated annealing.

For both of these methods, a cost function, $C(\cdot)$, is defined. This is a real-valued function defined on the parameter space such that its minimum provides the optimum solution. A neighbourhood structure is defined on the parameter-space, a starting point, x_0, in the parameter-space selected and the cost function is evaluated at x_0. For steepest ascent, a point x, is then selected in the neighbourhood of x_0. If $C(x) < C(x_0)$ then x_0 is set equal to x and the procedure repeated until it converges to a minimum for C. The problem with this approach is that the final value of x_0 is likely to be a local minimum only. In simulated annealing this is avoided by allowing, initially, values of x to be accepted which result in an increase in C. This enables the process to move away from the neighbourhood of a local minimum. In simulated annealing a lower cost solution will always be accepted. Whether a higher cost solution is accepted depends on the size of the increase in cost and a control variable designed to allow higher cost solutions to be accepted with high probability early in the process but with decreasing likelihood as the process continues, until eventually, none is accepted.

To implement each of these a cascade method was used. Cascade methods have previously been applied to image analysis by Jubb and Jennison (1991) and in a modified form by Hurn and Jennison (1995). They involve forming pixels into larger, non-overlapping 2×2 blocks and averaging the colours of the pixels to obtain a colouring for the coarser resolution image. This process is repeated and at each repetition the scale of the image-restoration problem is reduced by a factor of four as is the variance of the noise. The image restoration algorithm is first applied to a coarse-resolution image and the solution to this is used as a starting point for the problem at the preceding finer resolution, and so on, until the original problem has been solved.

Chapter 9

Applications of wavelets to standardisation

9.1 Introduction

In section 9.2 of this chapter, we review methods that aim to standardise in the wavelet domain and describe our own work in this area. In section 9.3 we describe a method that treats the wavelength shift as a deformation function and aims to construct it in the wavelet domain.

9.2 Standardisation in the wavelet domain

A few workers have attempted to standardise NIR spectrometers by first transforming spectra to the wavelet domain. The main disadvantage of working in the wavelet domain is that information relating to the continuity of the spectra in the wavelength domain may be lost in the wavelet representation. The exception to this is the wavelet approximation at resolution level 1. Since this is obtained from the signal by subtracting the level 1 detail which is often treated as noise, it may provide better information than the original signal. Walczak et al. (1997) suggested an ingenious method of standardisation using

the resolution level 1 wavelet decomposition. We also give details of a wavelength correction method that is similar to the procedure proposed by Shenk and Westerhaus, but using the level 1 wavelet approximation rather than the raw spectra.

When the signal is decomposed at levels greater than 1, the wavelength structure is lost and consequently it is difficult to correct for a wavelength shift. For the same reason we can no longer exploit the smoothness properties of the regression coefficients. Bayesian methods that work well in the wavelength domain are based on smoothing and shrinking regression coefficients. In the wavelet domain only shrinking is available and this on its own does not allow much improvement on standard methods.

9.2.1 Methods that standardise in the wavelet domain

Probably the most successful method so far published is one suggested by Walczak et al. (1997). In their experimental work, Walczak et al. (1997) use the 10 orthogonal wavelet bases from the Daubechies family. The method adjusts the slave spectra to resemble the master spectra using a standardisation set. Master spectra in the standardisation set are transformed to the wavelet domain at level 1 using one of the Daubechies wavelet bases. The slave spectra are also transformed in the same way using the first of the Daubechies wavelet bases. Each wavelet coefficient on the slave instrument is related to the corresponding coefficient on the master instrument by univariate linear regression and the residuals are calculated and stored. The process is repeated using the same wavelet basis for the master spectra but using each of the 10 Daubechies bases in turn for the slave spectra, giving 10 sets of residuals and 10 regression coefficients for each wavelet coefficient. Then separately for each coefficient, the Daubechies basis for which the residuals are smallest is selected. For each wavelet coefficient the selected Daubechies basis and the corresponding regression coefficient are stored. To standardise a spectrum measured on the

slave instrument each wavelet coefficient must be calculated separately using the appropriate Daubechies basis and the corresponding regression coefficient. The wavelet coefficient for the selected Daubechies basis is transformed using the corresponding regression coefficient. When all the wavelet coefficients have been found the inverse wavelet transform corresponding to the wavelet basis used to transform the master spectra is used to yield the standardised spectrum.

The use of different wavelet bases is intended to compensate for the wavelength shift, by weighting spectral values differently within a window. Walczak et al. (1997) report improvement over PDS using this method.

Park et al. (2001) compare several methods including PDS and DS and a method based on wavelet transforms. Spectra for master and slave instruments are transformed to the wavelet domain and PDS with window-size 1 (i.e. SW without the wavelength shift correction) is applied in the wavelet domain but only to the upper level coefficients. Wavelet coefficients below a certain level are left unchanged. This level is selected as the one that gives the best RMSEP on the standardisation set. Results reported are similar to those for PDS.

Yoon et al. (2002) use the universal threshold to compress the spectral data and DS on the reduced set of wavelet coefficients to standardise. This method required 10 or more standardisation samples to improve on PDS. Neither of these methods is likely to work well where there is a wavelength shift.

As mentioned previously, wavelet decomposition destroys the continuity of the spectra, however wavelet decomposition at level 1 separates the signal into an approximation that is a very slightly smoothed version of the original and level 1 detail many of whose coefficients are likely to be close to zero and can be treated as noise. We applied PDS with window-size 1 to the level 1 approximation of the master and slave spectra. We also applied PDS with Bayesian shrinkage. We decomposed the slave and master spectra completely in the wavelet domain and applied PDS to the wavelet coefficients. None of

Table 9.1: RMSEP using NIR data with SW applied to the wavelet approximation at level 1

Reference value	1	2	3	4	5
Standardisation set size 5					
RMSEP(25 spectra)	0.7492	0.2927	0.1083	0.1623	0.6368
RMSEP(30 spectra)	0.7685	0.2811	0.1051	0.1831	0.7031

these approaches improved on PDS.

One advantage of using wavelets is that one can exploit the thresholding techniques that are used for denoising. We used the thresholding subroutines from Matlab's wavelet toolbox to denoise the spectra before applying PDS with window 1. We also applied Bayesian shrinkage to the denoised spectra. None of these methods worked better than PDS alone.

An adaptation of the Shenk-Westerhaus method in the wavelet domain was a substantial improvement over the corresponding method in the wavelength domain. A wavelet transformation at resolution level 1 was applied to the master and slave spectra in the standardisation set. Wavelet 4 in the Daubechies family was used following the conclusions of Fearn and Davies (2003) though the choice was not critical. The level 1 approximation coefficients were then used to find the wavelength shift at each wavelength separately. The method is identical with the SW wavelength shift correction method. A first derivative treatment was applied to the wavelet coefficients. Then the correlation between master spectra for a fixed coefficient and slave spectra in the neighbourhood of that coefficient were compared. The point where correlation was greatest was used to define the shift. Once the slave spectra have been corrected for wavelet shift, the absorbance shift is corrected by regressing master absorbance on slave absorbance, wavelength by wavelength.

Results of standardisation for the NIR data using the above method are shown in Table 9.1

9.3 Estimating the wavelength shift as a deformation function

9.3.1 Introduction

Of the methods of standardisation where there is a wavelength shift, the method used by Shenk and Westerhaus to correct the wavelength shift does not appear to work well, while although PDS performs better, as mentioned previously, if our assumption that the difference between slave and master spectra is due to a wavelength shift or an absorbance shift or both, the solution does not reflect the perceived error. In an attempt to improve on these methods we adopt a model motivated by the Shenk-Westerhaus method but estimate the wavelength shift by treating it as a deformation between the slave spectra and the master spectra and using the ideas described in section 5 of the previous chapter. Two features of the wavelet representation of a function make it an attractive choice for representing the wavelength shift, f. Firstly, the wavelet coefficients can be assumed to be independent of each other and secondly, because there is insufficient data to determine f exactly, using wavelet thresholding should lead to a parsimonious representation that is an adequate approximation of f.

9.3.2 Model

Assuming, as we have done previously, a linear absorbance shift, our model is

$$X_m(i,j) \sim N(X_s(i, j + f(j))g(j) + h(j), \sigma^2)$$

where i indexes samples, $f(j)$ represents the wavelength shift at wavelength j and $h(j)$ and $g(j)$ are the regression coefficients for the absorbance shift at this wavelength. We assume that the wavelength shift is less than one unit - in the case of our data, 2 nanometers - so that $|f(j)| < 1$. We simplified the problem by restricting the value of $f(j)$, because for our data a greater wavelength shift

was unlikely, but it would be only a slight complication to allow shifts that were greater than 1 unit. In order to estimate X_s at non-integral points we use linear interpolation between neighbouring points:-

$$X_s(i, j + f(j)) = X_s(i, j) + |f(j)|(X_s(i, j + sign(f(j))) - X_s(i, j))$$

We use a wavelet basis to represent the function f

$$f(x) = u\phi(x) + \sum_{l,k} w_{lk}\psi_{lk}(x),$$

where $\phi(x)$ is the scaling function and $\psi_{lk}(x)$ is the wavelet function

$$\psi_{lk}(x) = \psi(2^l x - k).$$

Here we have allowed the constant $2^{l/2}$ to be absorbed into w_{lk} so the the ψ_{lk} are orthogonal but not orthonormal.

We place a prior distribution on the wavelet coefficients given by

$$w_{lk} \sim \pi_l \delta(0) + (1 - \pi_l)N(0, \tau_l^2)$$

Here $\delta(0)$ represents a point mass at 0 and π_l is the probability that a wavelet coefficient at resolution level l is zero. Non-zero wavelet coefficients at the same resolution level are assumed to be independently and identically distributed.

With appropriate choice of parameters this prior will discourage non-zero coefficients at higher levels of resolution, leading to a parsimonious representation of f in the wavelet domain. The priors on the wavelet coefficients that define f perform a function similar to the prior on the regression coefficients in PDS or S-W; they smooth f and shrink it.

9.3.3 Cost function

Our intention is to use simulated annealing and ICM to estimate f. In order to use these methods we must define a cost function. The cost function is based

on the Bayesian posterior for our model. The likelihood, $l(X_m, X_s|f, g, h, \sigma^2)$ is given by

$$l(X_m, X_s|f, g, h, \sigma^2) \propto \exp(-\frac{1}{2}(trace(X_m - X_{sfgh})^T(X_m - X_{sfgh})/\sigma^2))$$

where $X_{sfgh}(i, j) = X_s(i, j + f(j))g(j) + h(j)$ with g and h found by regression of the master spectra on the wavelength adjusted slave spectra.
The prior probability density function for w_{lk} is given by

$$\prod_{l,k}(\pi_l I(w_{lk} = 0) + (1 - \pi_l)\frac{1}{\sqrt{(2\pi\tau_l^2)}}\exp(-\frac{1}{2}\frac{w_{lk}^2}{\tau_l^2})I(w_{lk} \neq 0)).$$

As explained in the previous chapter, this prior reflects the fact that functions can often be represented parsimoniously in the wavelet domain, by setting coefficients to zero with probability π_l and shrinking non-zero coefficients towards zero. Consequently zero and small values of the wavelet coefficients for f will be favoured.

The posterior for f is proportional to

$$\exp(-\frac{1}{2}trace(X_m - X_{sfgh})^T(X_m - X_{sfgh})/\sigma^2))$$
$$\prod_{l,k}(\pi_l I(w_{lk} = 0) + \frac{(1 - \pi_l)}{\sqrt{(2\pi\tau_l^2)}}\exp(-\frac{1}{2}\frac{w_{lk}^2}{\tau_l^2})I(w_{lk} \neq 0)).$$

The posterior is maximised when the cost function,

$$trace(X_m - X_{sfgh})^T(X_m - X_{sfgh})$$
$$-2\sigma^2\sum_{lk}\log(\pi_l I(w_{lk} = 0) + \frac{(1 - \pi_l)}{\sqrt{(2\pi\tau_l^2)}}\exp(-\frac{1}{2}\frac{w_{lk}^2}{\tau_l^2})I(w_{lk} \neq 0)),$$

is a minimum. Here the second term in the cost function acts as a penalty for non-zero or large wavelet coefficients, with the model variance, σ^2, determining the weight given to this penalty.

Our aim is to find the set of coefficients w_{lk} that minimises this cost function.

9.3.4 Choice of wavelet

Our choice of wavelet was determined by ease and speed of computer implementation. Most wavelets are defined in terms of Fourier transforms while the Haar function is a simple step function for which efficient programs can easily be written. It was for this reason that we chose, at least initially, to use the Haar wavelet.

The Haar function is given by:-

$$\psi(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{otherwise} \end{cases}$$

This function is the mother wavelet in the Haar system. The scaling function takes the value 1 on the interval [0,1] and is zero elsewhere.

9.3.5 Choice of parameters

The parameters for which values are needed are π_l, the probability that a wavelet coefficient at resolution level l is zero, τ_l^2, the variance of non-zero wavelet coefficients at level l and σ^2, the model variance.

We need to consider the choice of π_l and τ_l^2 at the same time, since the relative sizes of the probabilities that a particular coefficient is zero or non-zero depend on both. The cost of a non-zero coefficient, w, is

$$-2\sigma^2 \log((1-\pi_l)\frac{1}{\sqrt{(2\pi\tau_l^2)}} \exp(-\frac{1}{2}w^2/\tau_l^2))$$

which takes its minimum value $-2\sigma^2 \log((1-\pi_l)\frac{1}{\sqrt{(2\pi\tau_l^2)}})$, when $w = 0$. In order that f varies smoothly we need to avoid having a large number of non-zero wavelet coefficients especially at higher levels. To achieve this we choose values of τ_l^2 and π_l such that the minimum cost of a non-zero coefficient is greater than the cost of a zero coefficient

i.e.

$$\pi_l > (1 - \pi_l)\frac{1}{\sqrt{(2\pi\tau_l^2)}} \tag{9.1}$$

for larger values of l .

In their work on thresholding using the universal threshold Donoho and Johnstone (1994) suggest keeping lower resolution level coefficients even when they do not pass the threshold. In order to apply this in our context we need the reverse of inequality (9.1) for the first few values of l.

Since it is assumed that $|f(j)| < 1$ for all j we also need to choose τ_l^2 to limit the size of f. We used $\tau_l^2 = 0.16$ for all values of l. If, following Abramovich et al. (1998), we assume that $2^{-l}\tau_l^2$ has the form $2^{-\alpha l}C_1$, then our assumption about τ_l^2 is equivalent to assuming $\alpha = 1$ and $C_1 = 0.16$. With $\pi_l = 1 - min(1, 2^{-bl}C_2)$, Abramovich et al. suggested values for b lying between 0 and 1: we use $b = 0.4$ and $C_2 = 2$.

The parameter σ^2 was chosen so that the error term and the penalty term in the cost function were of roughly the same order. We chose the number of iterations at each level to be large enough to ensure convergence by checking that up-dating of the cost function had ceased.

9.3.6 Implementation

Our problem is slightly different in emphasis from the deformation problems described in the previous chapter in that the aim is not to converge to a function, f, which minimises the squared error between master and slave on the standardisation set. We need a solution that provides a good fit for the entire dataset. Also because our final objective is prediction, the fit at some wavelengths is more critical than at others.

Because of the way wavelets are defined, we chose to use a wavelength range that was a power of two for each of our examples. For the NIR data the range was 1090 to 1600 nanometers, omitting part of the spectrum in which there was little information, for the CORN data, 1100 to 2122 nanometers.

f is estimated using two different methods:- steepest ascent or ICM described in section 8.5 and simulated annealing. For each of these methods f and hence all the wavelet coefficients are initially assumed to be zero. At each level wavelet coefficients are randomly sampled in the neighbourhood of their current values and if the sampled values are accepted they replace the current values in the expression for f. The cost of using those coefficients together with the best values already selected, is evaluated. This involves applying the wavelength correction determined by f to X_s to give X_{sf} and then regressing X_m on X_{sf} at each wavelength to find g and h. The process continues until convergence is reached.

The order in which the wavelet coefficients are determined is motivated by the 'cascade method' of Hurn and Jennison (1995) and exploits the hierarchical nature of the wavelet representation. f is initially estimated at the coarsest level of approximation. The best (minimum cost) estimate sampled at that level is used as a starting point for estimating coefficients at the coarsest level of detail. Again, the best estimates sampled at this level are assumed as a starting point for finding and fixing coefficients at finer levels of detail.

ICM and simulated annealing differ in the criteria for acceptance of sampled wavelet coefficients at a given level. For ICM new coefficients are accepted as long as the cost function calculated using them is lower. For simulated annealing there is the possibility, earlier on in the process, of wavelet coefficients that yield an increased cost being accepted. The probability of this happening is controlled by a temperature schedule that is defined so that this probability diminishes with time allowing increased cost solutions to be accepted at the beginning of the process but eventually allowing no higher-cost alternatives to be accepted.

ICM is comparatively quick to apply and will converge at least to a local minimum. Disadvantages are that it cannot escape a local minimum and is extremely sensitive to the starting configuration, (Hurn and Jennison (1995)).

For our application these disadvantages are not too serious. Our starting configuration reflects our prior expectation about the solution and since there is insufficient information to provide a unique solution, a local minimum may well be adequate. Repeated application gave results that were good but not consistent. In an attempt to improve consistency we used simulated annealing. We also tried estimating the wavelength shift as the mean of all the values of f found using a given set of parameters and method.

9.3.7 Results and discussion

For the NIR data and a standardisation set of size 5 we performed 40 runs with the same parameters. In table 9.2 we give the mean and standard deviation of prediction errors for each of the response variables over 40 runs. We also give the prediction errors when the wavelength shift is calculated as the mean of the 40 shifts found previously. In table 9.3 we give equivalent results when simulated annealing was used. We performed 20 runs with standardisation sets of size 8 and 10, with and without simulated annealing. Results of these are in tables 9.4, 9.5, 9.6 and 9.7.

The wavelength shifts found using 5, 8 and 10 standardisation samples and taking the mean of the wavelength shifts found over 40 runs using ICM are shown in figure 9.1. The differences between the three graphs are small especially in the wavelength range 1300 - 1600 nm. where most of the information lies, suggesting that there is little advantage in using larger standardisation sets to determine the wavelength shift. This view is reinforced in the tables where it can be seen that results for 5, 8 and 10 samples are fairly close, though, not surprisingly, using 5 samples offers more consistent results. Simulated annealing makes slight improvements when 5 or 8 samples are used for standardisation though a consequence of using this method is that results are more variable. Using these standardisation sets produces very good results in all cases. In particular, taking the mean of several values for the wavelength

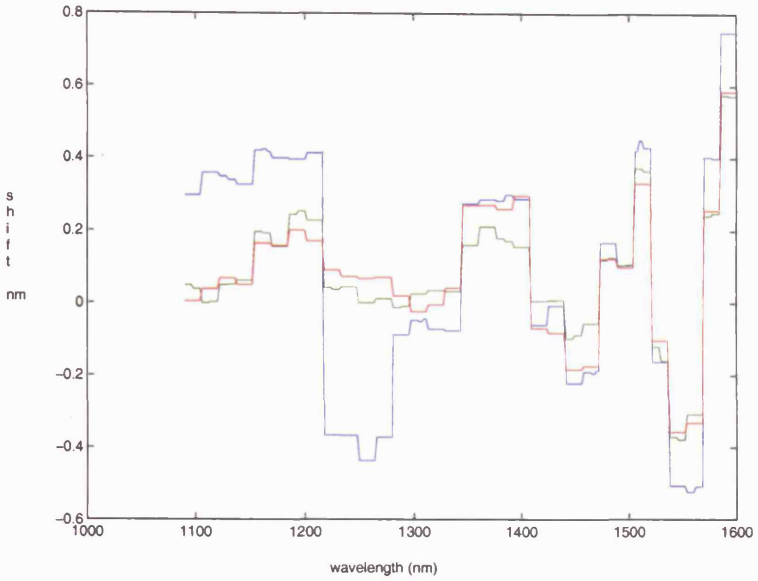


Figure 9.1: Graph showing the wavelength shift function found using standardisation sets of size 5 (blue), 8 (green) and 10 (red), using ICM.

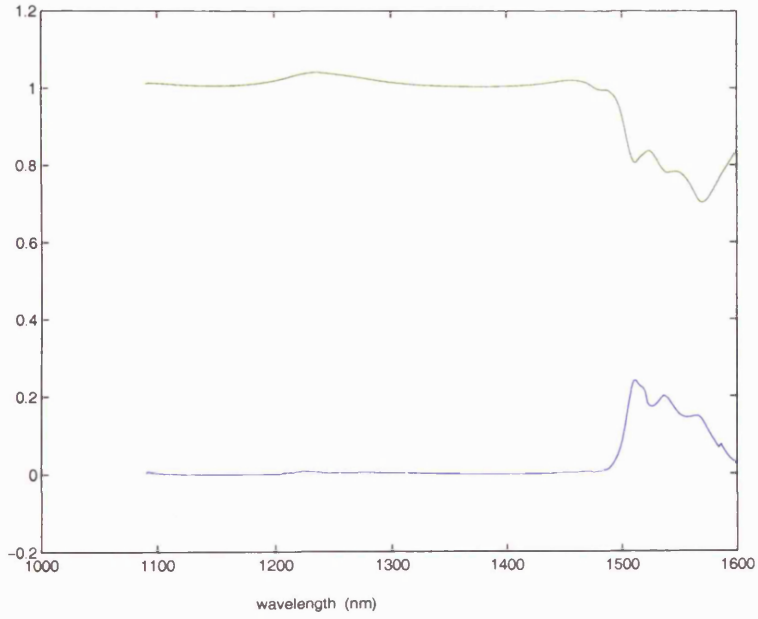


Figure 9.2: Regression coefficients for NIR data using 10 standardisation samples.

shift appears to work well. As one would expect, using larger standardisation sets gives slightly improved results.

The regression coefficients are shown in figure 9.2. The upper graph, the slope, is close to 1 while the intercept is very close to zero except in the wavelength range 1500 to 1600 nmThese graphs are remarkably smooth compared with those produced by other methods. (See, for example, 5.3(a) or 7.1). This suggests that the wavelength shift has been more accurately estimated here.

The same method was also used to standardise the corn data. There appears to be no wavelength shift in this data and this was confirmed by our results. The estimated size of the wavelength shift was always less than 0.1. As a result the procedure for finding the wavelength shift had the effect of smoothing the regression coefficients and the results were similar to the full Bayesian approach of Chapter 5 when there was no correction for wavelength shift. In table 9.8 we give the predictions from a single application of ICM and also those using the full Bayesian method assuming no wavelength shift.

NIR DATA

Table 9.2: Results for 40 applications of ICM using a standardisation set of size 5

Reference value	1	2	3	4	5
Results using 25 spectra not in standardisation set					
mean(RMSEP)	0.7055	0.2796	0.0946	0.1739	0.5633
SD(RMSEP)	0.0375	0.0088	0.0050	0.0137	0.0364
f taken as mean of 40 values from above results					
RMSEP	0.6889	0.2774	0.0931	0.1723	0.5501
Results using all 30 spectra					
mean(RMSEP)	0.7451	0.2788	0.0911	0.1891	0.6090
SD(RMSEP)	0.0399	0.0094	0.0052	0.0133	0.0392
f taken as mean of 40 values from above results					
RMSEP	0.7311	0.2768	0.0896	0.1877	0.5982

Table 9.3: Results for 40 applications using simulated annealing using a standardisation set of size 5

Reference value	1	2	3	4	5
Results using 25 spectra not in standardisation set					
mean(RMSEP)	0.7064	0.2838	0.1030	0.1832	0.5653
SD(RMSEP)	0.0540	0.0190	0.0165	0.0280	0.0526
f taken as mean of 40 values from above results					
RMSEP	0.6742	0.2716	0.0986	0.1785	0.5395
Results using all 30 spectra					
mean(RMSEP)	0.7442	0.2822	0.0990	0.1974	0.6081
SD(RMSEP)	0.0549	0.0169	0.0157	0.0259	0.0545
f taken as mean of 40 values from above results					
RMSEP	0.7163	0.2711	0.0947	0.1932	0.5862

NIR DATA

Table 9.4: Results for 20 applications of ICM using a standardisation set of size 8

Reference value	1	2	3	4	5
Results using 22 spectra not in standardisation set					
mean(RMSEP)	0.6795	0.3072	0.1095	0.2087	0.5452
SD(RMSEP)	0.0555	0.0325	0.0140	0.0338	0.0503
f taken as mean of 20 values from above results					
RMSEP	0.6500	0.3380	0.1141	0.2235	0.5157
Results using all 30 spectra					
mean(RMSEP)	0.7351	0.3123	0.1063	0.2188	0.5931
SD(RMSEP)	0.0490	0.0345	0.0144	0.0301	0.0488
f taken as mean of 20 values from above results					
RMSEP	0.7147	0.3463	0.1117	0.2302	0.5610

Table 9.5: Results for 20 applications using simulated annealing using a standardisation set of size 8

Reference value	1	2	3	4	5
Results using 22 spectra not in standardisation set					
mean(RMSEP)	0.6772	0.3043	0.1115	0.2093	0.5490
SD(RMSEP)	0.0579	0.0312	0.0161	0.0350	0.0495
f taken as mean of 20 values from above results					
RMSEP	0.6421	0.3346	0.1157	0.2233	0.5148
Results using all 30 spectra					
mean(RMSEP)	0.7336	0.3084	0.1083	0.2194	0.5959
SD(RMSEP)	0.0497	0.0314	0.0166	0.0312	0.0482
f taken as mean of 20 values from above results					
RMSEP	0.7082	0.3417	0.1133	0.2300	0.5596

NIR DATA

Table 9.6: Results for 20 applications of ICM using a standardisation set of size 10

Reference value	1	2	3	4	5
Results using 20 spectra not in standardisation set					
mean(RMSEP)	0.6656	0.2786	0.1047	0.1978	0.5368
SD(RMSEP)	0.0642	0.0167	0.0106	0.0246	0.0547
f taken as mean of 20 values from above results					
RMSEP	0.6151	0.2981	0.1062	0.2086	0.4946
Results using all 30 spectra					
mean(RMSEP)	0.7243	0.2895	0.1014	0.2094	0.5897
SD(RMSEP)	0.0538	0.0141	0.0104	0.0219	0.0510
f taken as mean of 20 values from above results					
RMSEP	0.6804	0.3155	0.1048	0.2203	0.5454

Table 9.7: Results for 20 applications using simulated annealing using a standardisation set of size 10

Reference value	1	2	3	4	5
Results using 20 spectra not in standardisation set					
mean(RMSEP)	0.6728	0.2696	0.1010	0.1866	0.5430
SD(RMSEP)	0.0630	0.0232	0.0103	0.0246	0.0540
f taken as mean of 20 values from above results					
RMSEP	0.6131	0.2834	0.1020	0.1959	0.4967
Results using all 30 spectra					
mean(RMSEP)	0.7319	0.2833	0.0982	0.2008	0.5949
SD(RMSEP)	0.0539	0.0139	0.0098	0.0217	0.0512
f taken as mean of 20 values from above results					
RMSEP	0.6798	0.3036	0.1011	0.2100	0.5469

CORN data

Table 9.8: ICM with a standardisation set of size 5 and Bayes with no wavelength correction

Reference value	1	2	3	4
RMSEP	0.3192	0.1032	0.1590	0.3599
No wavelength shift	0.3193	0.1032	0.1590	0.3598

Chapter 10

Robust methods

Introduction[1]

The standardisation methods which we have so far discussed have all been based on adjusting the slave spectra so that they resemble the corresponding spectra on the master instrument. The emphasis has been on standardising instruments where there is a wavelength shift. The methods discussed in this chapter produce calibrations which are robust to several different instruments. They are unlikely to work well if the spectral difference involves a wavelength shift.

10.1 The repeatability file

Possibly the simplest method of producing a calibration that is robust to different instruments is to add spectra from those instruments and their reference values to the calibration set for the master instrument. This approach has been tested by various workers, see for example Hardy et al. (1996) but without a large number of extra spectra for each instrument, this method is unlikely to

[1] This chapter is based on "Transfer by orthogonal projection", Andrew and Fearn, (2004)

...

work well.

Westerhaus (1991) suggested a modification of the above method which was found to produce improved calibrations. See for example Shenk and Westerhaus (1991), Tillmann et al. (2000). For this, spectra and reference values from the same samples on the master instrument and each of the instruments to be standardised are required. For each sample separately, spectra are mean-centred by subtracting the mean of all the spectra for that sample from each spectrum. Since the spectra are produced from the same sample they will have the same reference value and consequently the mean-centred spectra will have zero reference value. The repeatability file is formed from the mean-centred spectra. Members of the repeatability file are given zero reference values and added to the calibration set. The calibration model is,

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ kQ \end{pmatrix} \beta + \epsilon.$$

Here y is the vector of reference values for the spectra X for the master calibration set, Q is the matrix of spectra from the repeatability file and 0 is a vector of zeros, representing the reference values for the repeatability file. k is a scalar multiple which can be chosen to give the repeatability file an appropriate weight.

Algebraically this set is almost identical with one formed by simply adding the extra samples to the calibration set, so that if LSR were used and $k = 1$ the results would be the same. Because it is necessary to use some form of regularisation results differ. If PCR is used results may be worse. This is because PCR constructs factors that are linear combinations of the original spectra that maximise variance. If the between-instrument variation is large PCR factors will reflect this variation and consequently will be useless for prediction. Because PLS factors are correlated with the reference values, using PLS with a repeatability file will tend to result in factors for which the repeatability file contributions are close to zero. Using PLS, improved calibrations have been

found.

Westerhaus noted that using the repeatability file is related to generalised ridge regression. This can be seen by considering the LSR solution to the above regression model:-

$$\beta = (X^T X + k^2 Q^T Q)^{-1} X^T y.$$

In ridge regression the matrix $Q^T Q$ is the identity matrix and k, the shrinkage factor for β, is usually small. The result is that each of the entries in β is shrunk towards zero. The difference here is that Q is not full rank and consequently shrinkage occurs only in the directions defined by the repeatability file - the directions in which between-instrument variation occurs. From a Bayesian perspective this solution is equivalent to placing vague priors on all components of β except those corresponding to the directions in which the between-instrument variation occurs. In these directions the shrinkage factor, k^2, is the ratio of the model variance to the prior variance of the regression coefficient.

Given this analysis, allowing k to become large so that between-instrument variation is shrunk to zero would seem a sensible approach. This might work in LSR if a LSR solution existed. Using PLS, the expression for the regression coefficients has k^2 in the denominator so that as $k \to \infty$, $\beta \to 0$. This can be proved using the Krylov basis for the PLS factors (Helland (1988)).

Only using PCR is there a possibility of a solution with k large. The effect of using PCR with large k is to enlarge the between-instrument variation, so that the the first few scores found by PCR will reflect between-instrument variation only. The remaining scores will be useful for calibration. The regression coefficients corresponding to the first few scores will be small so that the between-instrument contribution will be unimportant compared with the contribution of the relevant scores. A proof that for large k PCR scores are separated into scores for Q and scores orthogonal to Q is given in Appendix A.

10.2 Transfer by orthogonal projection

One disadvantage of using a repeatability file is, as we have seen, that it introduces spectral data that may affect adversely the formation of factors in the regularisation step. Transfer by orthogonal projection, (TOP), the method that we have developed, projects the spectra onto a subspace orthogonal to the subspace in which most of the between-instrument variation occurs, so that factors found in the regularisation step as well as calibrations developed from them are robust to between-instrument variation.

10.2.1 Method

We first form the subspace which describes between-instrument variation. We assume that, as for the repeatability file, we have spectra for a few different samples produced on each of m instruments, the master instrument and each of the instruments we intend to standardise. For each instrument separately we average over the spectra for the given samples giving an $m \times w$ matrix. We perform PCA on the $m \times w$ matrix of average spectra. This will give $m - 1$ principal components that describe the between-instrument variation. We select enough PCs to capture most of the variation. The matrix, P, formed from these PCs will, because of the way PCA loadings are constructed (see 2.4.2), have orthonormal columns. We project the $n \times w$ matrix, X, of spectra on the master instrument onto the orthogonal complement of P, giving,

$$\tilde{X} = X(I - PP^T)$$

Finally we calibrate the instruments by regressing y on $\tilde{X}$. Because of the method of construction of $\tilde{X}$, the calibration will be robust to the between-instrument variation and consequently work equally well on all instruments involved in the process.

The choice of the number of PCs used to define the subspace of between-instrument variation is clearly important, but the only reliable method of mak-

ing this choice is selecting those PCs for which the RMSEP for the different instruments is minimised. Roger et al. (2003) describe a method similar to the one described here, which they apply to the elimination of variation caused by temperature differences in fruit. They suggest a method of selecting PCs based on Wilks' Λ. Since the normality conditions which motivate the distributional assumptions about Wilks' Λ are unlikely to be met, a significance test based on this statistic is invalid and any decision based on it is bound to be subjective.

Our method depends on selecting enough PCs to capture most of the between-instrument variation which is again subjective. Consideration of the RMSEP suggested selecting too few PCs was more of a danger than selecting too many.

10.2.2 Comment

There are obvious similarities between TOP and orthogonal signal correction (OSC), described in section 3.3.1. Both remove dimensions from the spectral data before calibration. The difference is that TOP uses the extra information from the standardisation set to remove exactly those dimensions that interfere with the transfer. OSC just removes dimensions orthogonal to y - those that are irrelevant to calibration.

10.3 Relation between repfile and TOP

As indicated in section 10.1 for calibration using MLR the repeatability file method is equivalent to shrinking, with shrinkage factor k, components of β that act on the subspace spanned by the repeatability file, Q. As $k \to \infty$ these tend to zero so that Q is mapped to zero. This is equivalent to regressing y on the orthogonal complement of Q, which is exactly what TOP does. The equivalence of TOP and repeatability file with $k = \infty$ is proved in Appendix B in the case of a single sample standardisation set. The result is true for larger

sets as long as the repeatability file spans the same subspace as P, the space of between-instrument variation generated by TOP. This will be the case if the difference between spectral responses of the different instruments for a given sample is an absorbance shift that is independent of the sample.

10.4 Experimental details

We applied both TOP and the repeatablity file method to two of our datasets, CORN and BARLEY.

10.4.1 Standardisation sets

We used a set of 5 samples and also a single sample. The set of five was selected as previously described. The single sample was selected following Shenk and Westerhaus (1991) to be the sample closest in Euclidean distance to the mean of all the samples for the master instrument. As for previous methods the samples in the standardisation set were included (measured on the master instrument only) in the calibration set, but removed from the assessment of transfer to other instruments.

10.4.2 Data treatment and results

A first derivative treatment was applied to the spectra of the barley data. The corn data were analysed both with and without a first derivative treatment. All samples on one instrument, an arbitrarily chosen master, were used to calibrate that instrument. PLS and PCR were used with cross-validation to decide the number of factors: 11 factors for the barley data and 4 for the corn data with PLS and 7 with PCR using either raw or first derivative spectra. The same calibration was then used on the other instruments in the set with no adjustment. In applying TOP to the barley data, we used $d = 5$, accounting for 98% of the variation between instruments, though d=3, 4 or 6 gave very

Table 10.1: Root mean square errors of calibration and prediction for barley data using PLS

	RMSEC	RMSEP					
Instrument	1	2	3	4	5	6	7
Unadjusted	0.6412	0.9617	1.9962	0.7685	0.5116	1.1957	2.4209
Repfile	0.5667	0.8500	0.8074	0.6425	0.7931	0.6596	0.6551
TOP	0.6471	0.5091	0.5900	0.4377	0.4777	0.5581	0.6823

Table 10.2: Root mean square errors of calibration and prediction for barley data and PCR

	RMSEC	RMSEP					
Instrument	1	2	3	4	5	6	7
Unadjusted	0.7500	0.7352	2.0322	0.7257	0.9192	3.8542	1.5916
Repfile	1.1560	1.0869	1.9068	1.1815	1.1601	1.1740	1.6725
TOP	0.8875	0.7127	0.6607	0.6262	0.6624	0.6375	0.6952

similar results. For the corn data we obtained good results with d=1 or 2. The 1-dimensional space accounted for 86% of the between-instrument variation (99% without first derivative treatment), while since $m-1=2$, taking d equal to 2 meant that the all the estimated variation was removed. The results quoted here are for $d=2$. Tables 10.1 and 10.2 give RMSEC and RMSEP for the seven instruments in the barley example, whilst tables 10.3 and 10.4 give RMSEC and RMSEP for the three instruments in the corn example, for the first reference value (concentration of moisture). (Results using the other reference value were similar)

Both the repeatability file and TOP improve the transferability for the barley data. As expected, the repeatability file does not combine well with PCR, working much better with PLS. Again as would be expected, TOP works more or less equally well with either calibration method. In the corn example, the repeatability file works with first derivative but does not improve transferability

Table 10.3: Root mean square errors of calibration and prediction for corn data using raw spectra

	PLS			PCR		
	RMSEC	RMSEP		RMSEC	RMSEP	
Instrument	1	2	3	1	2	3
Unadjusted	0.0674	1.5430	1.7731	0.0722	1.3958	1.4736
Repfile	0.4154	1.9386	2.2075	0.4243	1.4227	1.2666
TOP	0.1081	0.2178	0.2135	0.1096	0.2026	0.2006

Table 10.4: Root mean square errors of calibration and prediction for corn data with first derivative treatment

	PLS			PCR		
	RMSEC	RMSEP		RMSEC	RMSEP	
Instrument	1	2	3	1	2	3
Unadjusted	0.0688	1.3869	1.7046	0.0746	1.3404	1.7456
Repfile	0.3029	0.8680	0.7393	0.3384	0.8004	0.4971
TOP	0.1132	0.1934	0.1966	0.1726	0.2067	0.2091

with raw spectra. The failure of the repeatability file with the raw corn data appears to be due to the fact that most of the difference between the spectra on the different instruments is due to a constant absorbance shift. This results in constant entries in the repeatability file leading to a calibration which is unchanged by the addition of the repeatability file. Contrary to what was predicted the repeatability file works well with PCR on the corn data after a first derivative treatment has been applied. This is due to the fact that the between-instrument variation is larger than, and not highly correlated with the between-sample variation and scores describing between-instrument and between-sample variation occur separately. The pretreatment makes little difference to the performance of TOP. In the cases where the repeatability file is effective, using PLS for the barley and first derivative spectra for the corn, TOP appears to work as well or better.

Some idea of the effect of TOP on the spectra may be obtained from Figure 10.1. This relates to the corn example without derivative treatment, and shows difference spectra for an arbitrarily chosen sample not in the standardisation set. The two spectra at the bottom of the plot are differences between instruments 2 and 1 and between instruments 3 and 1, before removal of between-instrument variation via TOP. The two spectra at the top of the plot are the corresponding difference spectra after TOP. One can see that instruments 2 and 3 still agree better with each other than with instrument 1, but that all the differences are reduced by an order of magnitude.

10.5 Calibration transfer to unseen instruments

One of the attractions of developing robust calibrations, compared with methods that adjust spectra, is that it should in principle be possible to transfer the calibration to further, as yet unseen, instruments of the same type. The success of this will depend of course on how representative of general between-instrument variability are the instruments represented in the standardisation

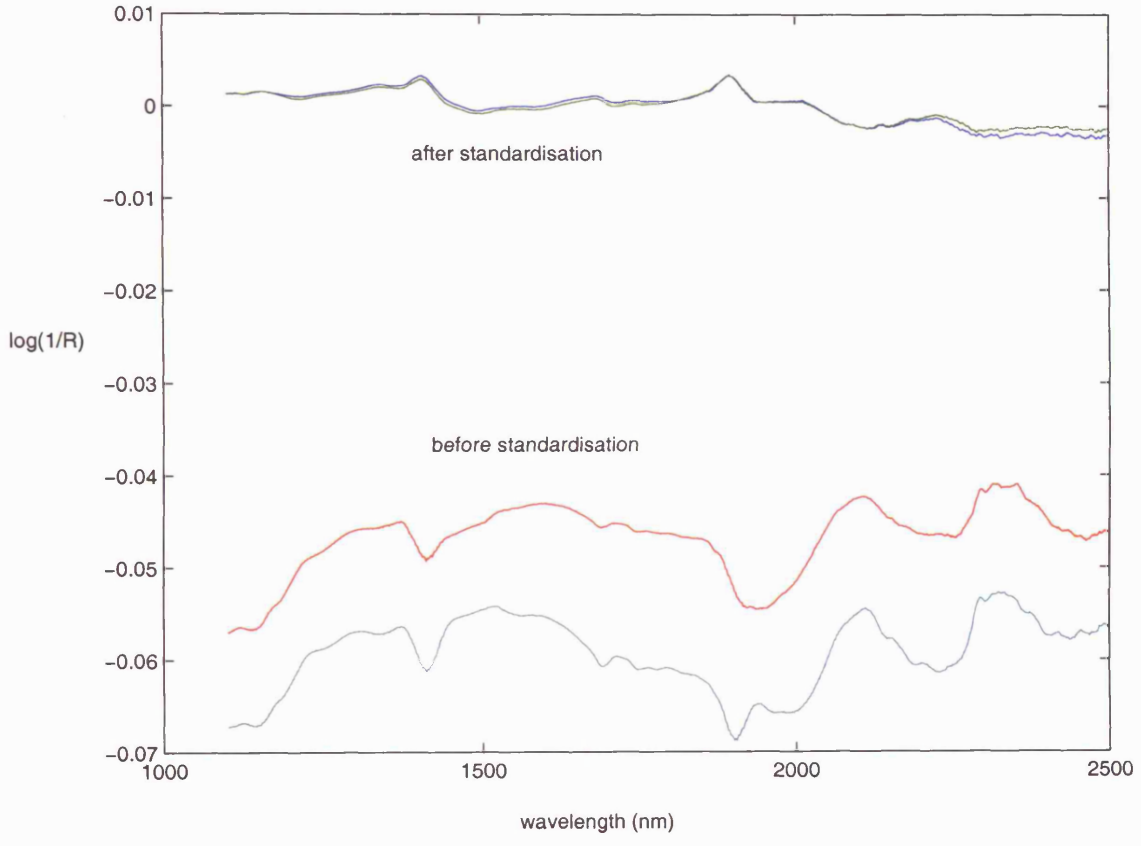


Figure 10.1: Difference spectra, instrument 2 minus instrument 1 and instrument 3 minus instrument 1, before and after standardisation using TOP, no derivative treatment

Table 10.5: RMSEPs for barley data, one instrument omitted from the standardisation set

Instrument	RMSEP on instrument					
omitted	2	3	4	5	6	7
2	0.6627	0.6496	0.5841	0.6367	0.6378	0.8536
3	0.6849	0.6453	0.6030	0.6689	0.6666	0.9491
4	0.6627	0.6758	0.5927	0.6393	0.6454	0.6767
5	0.6512	0.6581	0.5762	0.6213	0.6337	0.8078
6	0.6572	0.6332	0.5759	0.6377	0.7022	0.6788
7	0.6922	0.6606	0.5801	0.6526	0.6473	0.6710

Table 10.6: RMSEPs for corn data, one instrument omitted from the standardisation set

Instrument	RMSEP on instrument	
omitted	2	3
2	0.2096	0.2207
3	0.2126	0.2169

set.

We investigated, using the two examples, the transfer of a calibration to an unseen instrument using TOP. The procedure was to omit one instrument, base the standardisation set on the remaining ones, and then assess the resulting calibration on *all* the slave instruments, those involved in the calibration procedure and the one whose spectra had not been used at all. As for the previous investigations, instrument 1 was the master throughout. PLS was used for calibration, with 11 factors for the barley data and 4 for the corn data with first derivative treatment. The resulting RMSEPs are in table 10.5 for the barley data and table 10.6 for the corn. They result from the use of a single sample standardisation set. Similar results were obtained with five samples.

As can be seen from the RMSEPs on the diagonals of these two tables, the calibrations transfer successfully to instruments not included in the standardisation set in both of these examples.

Chapter 11

Summary and discussion

11.1 Introduction

Most of the material in this thesis has been based on two widely used methods for the standardisation of NIR spectrometers, the Shenk-Westerhaus (SW) patented method and Piecewise Direct Standardisation (PDS). Each of these methods reconstructs the absorbance at a particular wavelength on the master instrument from the absorbances on the corresponding slave spectrum, wavelength by wavelength. SW corrects for the two known sources of error separately. The wavelength shift is corrected first and after this, the absorbance shift. In PDS a less direct approach is used, making the assumption that the information required to determine the absorbance at a given wavelength on the master instrument from the corresponding slave spectrum is contained in the absorbances at that wavelength and those adjacent to it. By recontructing each absorbance independently of those adjacent to it, useful information is ignored.

Our methods have been designed to exploit more of the available information. The information that we use is that the wavelength shift between the slave instrument and the master instrument will vary smoothly from wavelength to wavelength and similarly absorbance shifts at adjacent wavelengths

are likely to vary smoothly. We also use the fact that the difference between slave and master spectra is likely to be small. All our methods work much better than SW or PDS alone, leading to substantial reductions in RMSEP.

We have based our models on the SW and PDS models, which can be summarised as

$$X_m - X_s B - 1_n \alpha^T \sim \mathcal{N}(I_n, \Gamma)$$

where X_m represents the matrix of absorbances for the master instrument and X_s that for the slave instrument. The matrix B is highly structured. For SW B is diagonal, while for PDS non-zero values occur only on or near the diagonal.

In a Bayesian approach, we placed priors on the regression coefficients B and α which reflect the relationship between adjacent coefficients. For each model it was necessary to specify or estimate parameters. We used two methods to do this. Firstly we selected values that minimised the RMSEP on the standardisation set. Secondly we placed priors on the parameters, leading to a hierarchical model and used MCMC to determine the joint distribution of the parameters given the data.

Dynamic Linear Modelling,(DLM) an alternative method uses the same models - SW and PDS - and places the same prior mean and correlation structure on the regression coefficients, but uses a recursive technique to establish their posterior distributions. This method is much faster than the Bayesian method.

The main problem with SW, at least with our data, was the failure of the wavelength correction step. We attampted to improve this by modelling the wavelength shift function in the wavelet domain. In a Bayesian approach we used wavelet thresholding to give an efficient representation of the wavelength shift function.

In our final method in chapter 10 we adopt a very different approach related to a method proposed by Westerhaus (1991) using a repeatability file. We

PDS	Bayes using MCMC	
PDS	DLM	
	Wavelength correction	Absorbance correctio
SW	maximising correlation between spectra	Bayes with plug-in p
SW	maximising correlation between spectra	Bayes using MCMC
SW	maximising correlation between spectra	DLM
SW	modelled in the wavelet domain	Linear regression

describe a method known as transfer by orthogonal projection describ Andrew and Fearn (2004). Both methods are designed to standardise se instruments simultaneously, by finding a robust calibration.

We now compare these methods. The main criteria used are speed of ation and performance, measured in terms of RMSEP on the slave instru We also, in section 11.3, compare the wavelength shifts when these are able.

11.2 Performance, parameter selection and sp of the methods

11.2.1 Parameter selection

Each of the standardisation methods involves parameters for which v must be selected. Some, such as the parameters for the Bayesian me using MCMC, do not need to be set very precisely and can be selecte observation from output graphs or, as in the wavelet method, from theore considerations. For the Bayes method using plug-in parameters, as expla

in chapter 5, Σ can be estimated from the data, but τ was estimated by selecting from a sequence of values the value that gave a minimum RMSEP on the standardisation set, or if the standardisation set was small, by minimising the RMSEP on a second set of samples, typical of the samples for which the calibration is to be used. The two parameters for DLM were selected by performing a sequence of runs with different combinations of the two parameter values and selecting values that minimised the RMSEP on the standardisation set.

11.2.2 Speed of methods

A run of DLM took 1.79 seconds for the NIR data so that although there were two parameters to estimate this could be completed rapidly. A single run of the Bayesian method took 11 seconds for SW and 45 seconds for PDS on the NIR data. The parameter τ for these was selected using a sequence of runs with different parameter values. MCMC with 2000 iterations on NIR data took 25 hours, so in spite of the fact that all parameters are estimated within the program MCMC was very time-consuming. The time required varies with the cube of the number of wavelengths so the CORN data would have taken several days using MCMC if the entire wavelength range were used. The wavelet method took 1.8 minutes per estimation. We used the average of 40 estimates which took 72 minutes.

11.2.3 Performance

We compare the RMSEP for each of the methods, considering standardisation using sets of sizes 5 and 10 and both datasets. Because there is little or no wavelength shift between slave and master on the CORN data, methods involving SW usually work as well as methods based on the PDS model, while for the NIR dataset methods using SW, since they rely on the SW wavelength adjustment, perform less well. Methods using the PDS model are the

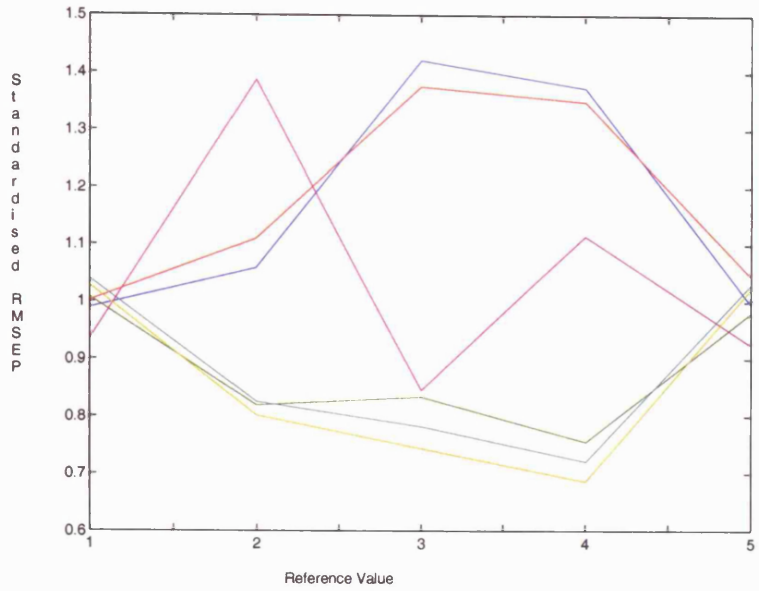
most successful for both datasets and both sizes of standardisation set. Using MCMC for SW or PDS is slightly better than using plug-in parameters. DLM worked well, though selecting parameters, especially for the NIR data, was slightly unsatisfactory. The wavelet method worked well with the NIR data and in fact with only 5 standardisation samples was the best method. Because the CORN data had virtually no wavelength shift the wavelet method was effectively the same as other methods for this dataset. Using 10 rather than 5 standardisation samples improved results considerably for the NIR data but made little difference for the CORN data.

In figure 11.1 we plot RMSEP for each of the reference values for the six main methods using a standardisation set of size 10: the PDS and SW models using MCMC and plug-in parameters, the wavelet method using ICM, and DLM. In figure 11.1a are shown RMSEP for the NIR data and in figure 11.1b, those for the CORN data. In figure 11.2 the RMSEP for each reference value for the same six methods using a standardisation set of size 5. For each reference value the RMSEP were standardised, by dividing each one by the mean of the RMSEP for that reference value.

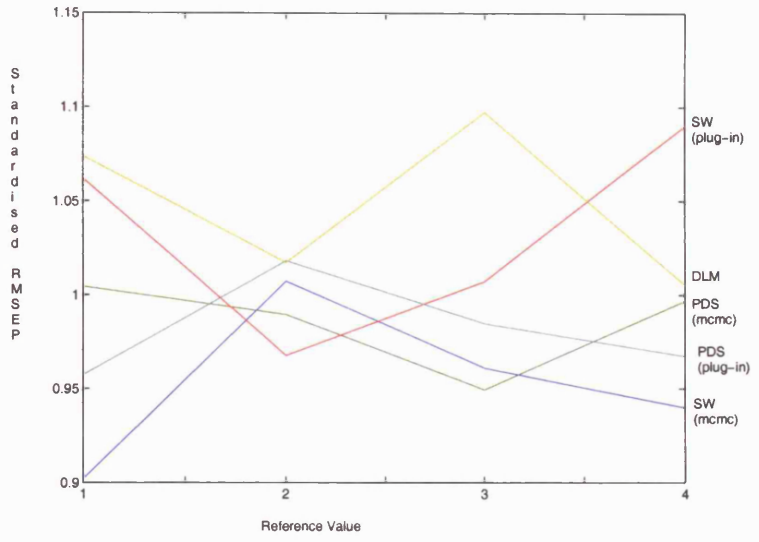
TOP described in chapter 10 is not strictly comparable with these methods because it can only be used when there is no wavelength shift, but where there is no wavelength shift, as in the CORN data, then it appears to work very well, producing RMSEP that are smaller than those for any of the other methods.

11.3 Comparison of wavelength shift functions

SW and the wavelet method described in chapter 9 are the only two methods that explicitly find the wavelength shift. Figure 11.3 shows the wavelength shift functions for the the wavelet method and for the SW method with Gaussian smoothing of the correlation coefficients used to estimate the wavelength shift, using a window of size 41. For both a standardisation set of size 10 was used. Allowing for the different methods of construction, the two graphs are very

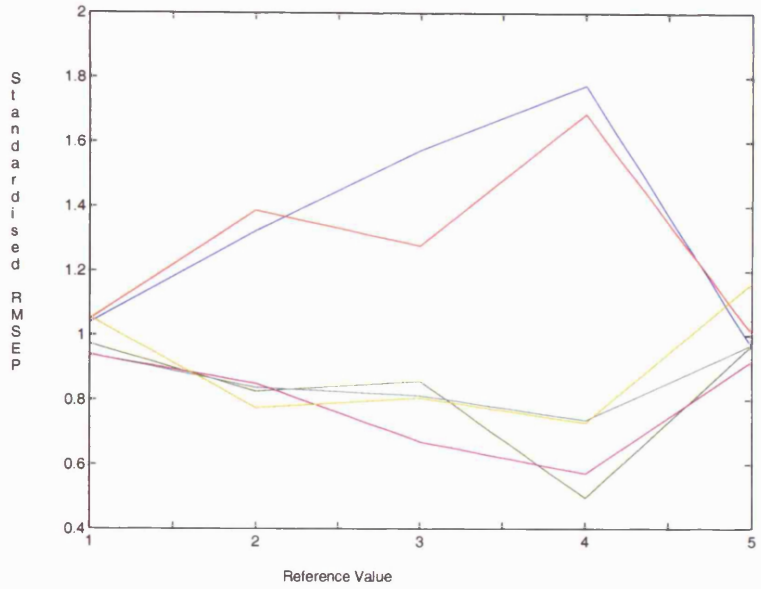


(a)

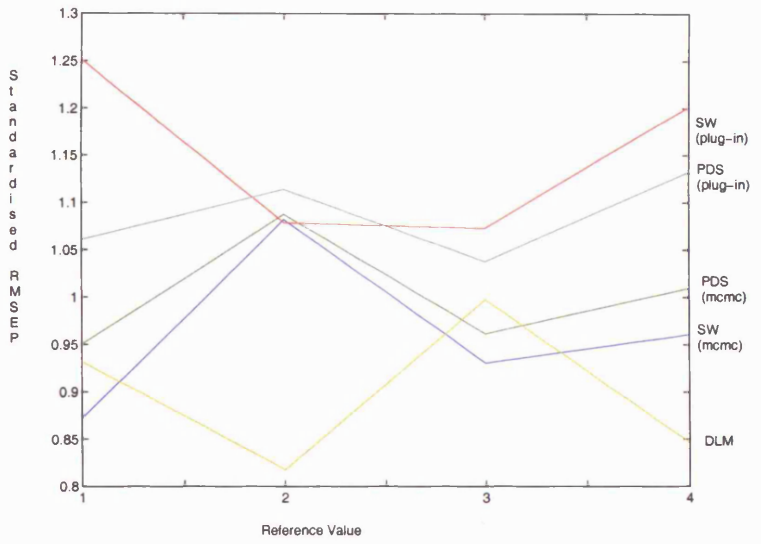


(b)

Figure 11.1: Standardised RMSEP for each reference value using 10 standardisation samples for PDS using MCMC(green), plug-in parameters (cyan), dlm (yellow), SW using MCMC (blue), plug-in parameters (red), wavelets (magenta), a) NIR data, b) CORN data



(a)



(b)

Figure 11.2: RMSEP for each reference value with 5 standardisation samples for PDS using MCMC (green), PDS using plug-in parameters (cyan), DLM (yellow), SW using MCMC (blue), SW using plug-in parameters (red), wavelets (magenta), a) NIR data, b) CORN data

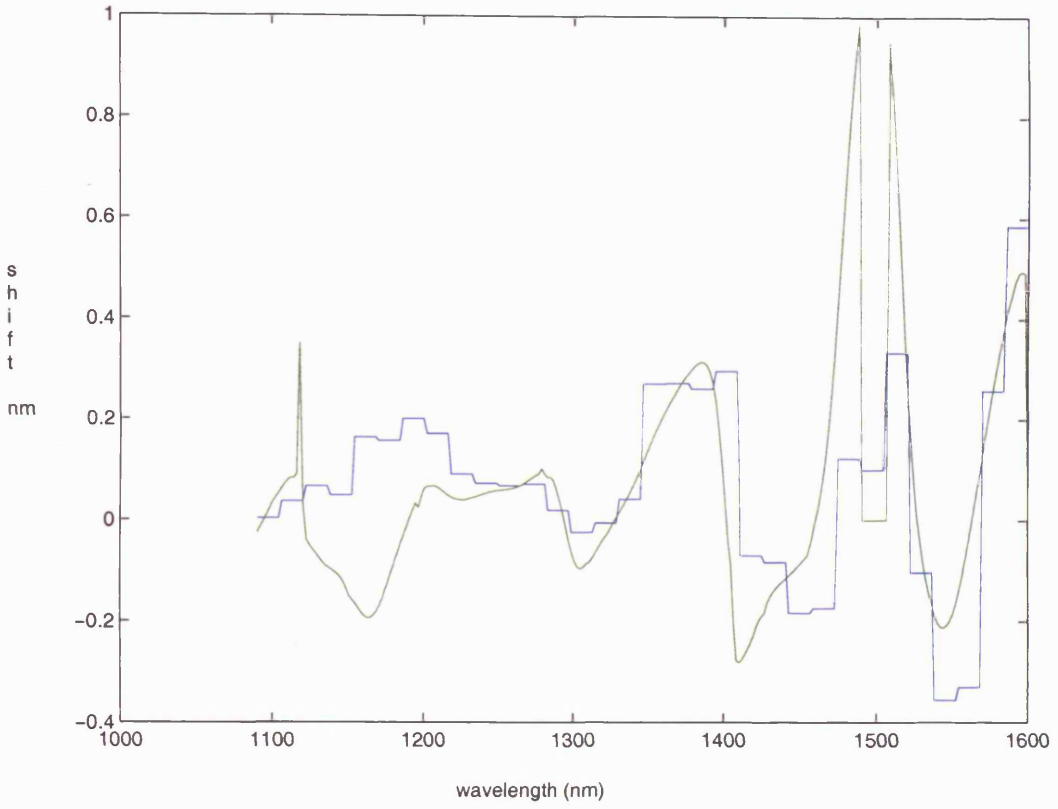


Figure 11.3: Wavelength shift functions found using wavelets (blue) and SW with Gaussian smoothing (green).

similar, the main difference being the sharp peaks occurring at about 1500 nm. in the SW graph, where the spectra are changing rapidly.

11.4 Comparison of the Bayesian method with dynamic linear modelling

Both these methods are based on the same model and both use Bayesian shrinkage and smoothing to adjust the regression coefficients. For each of these methods the model assumes that the product of the slave spectra and the regression coefficient, $X_S\beta$, gives an unbiased estimate of the corresponding master spectra, X_M. The correlation structure is the same and we used similar

values for the correlation coefficient between adjacent regression coefficients. It is interesting to compare the graphs of regression coefficients. For example the regression coefficients for the NIR data using the Bayesian method and plug-in parameters are shown in figure 11.4a and the regression coefficients for the same data using DLM in figure 11.4b. Because the data were not mean-centred for DLM the intercept is different, but the other coefficients are similar and in fact the intercepts, if the correction for mean-centring is taken into account are also similar.

Comparing the variances of the posterior distributions of the regression coefficients, those for the Bayesian method using MCMC (figure 5.4 b) are similar to those for the Kalman smoother. (figure 7.2). The greatest difference is in the 800 - 1000 nm. region where there is little information.

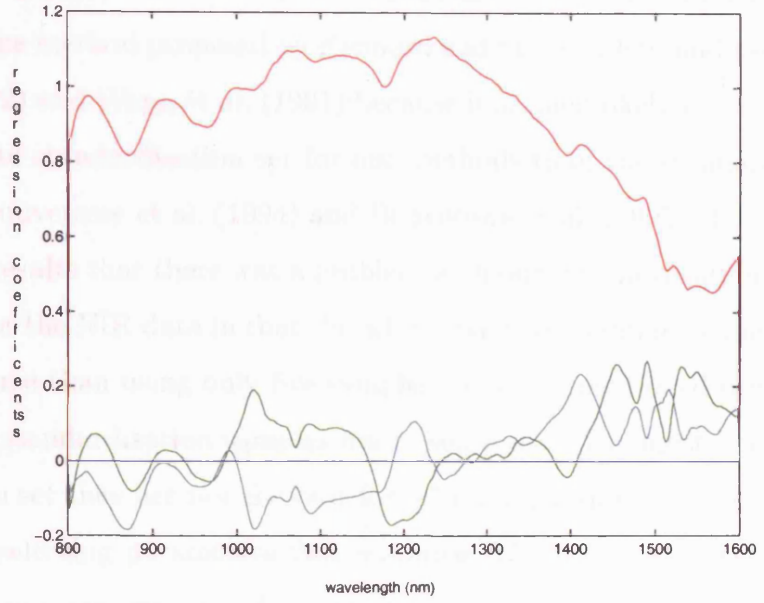
Using Bayes with MCMC has the advantage that parameters for variances were given priors and determined within MCMC, rather than, as in DLM, using plug-in values or relying on minimising the RMSEP on the standardisation set.

11.5 Window size

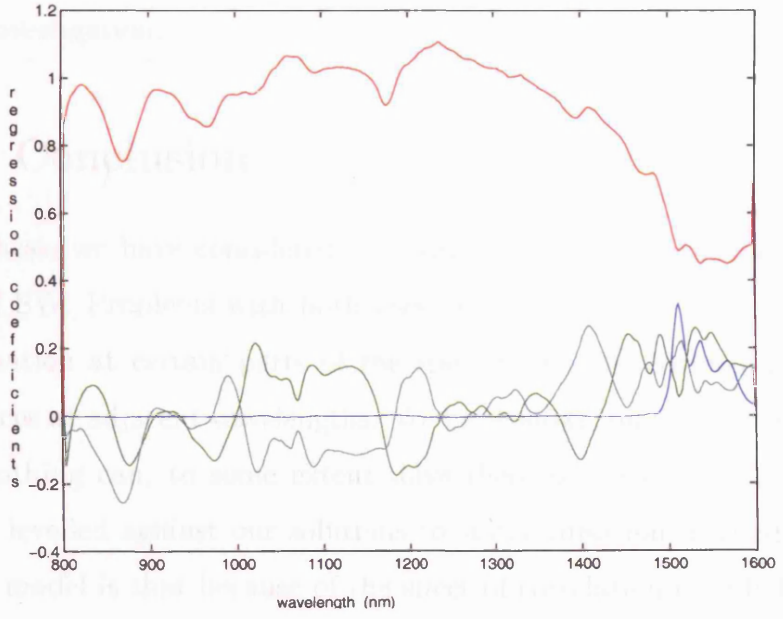
We chose to restrict the window-size to one that was realistic. It is unlikely that a NIR spectrometer would have a wavelength shift of more than 2 nanometers, so a window of size three was used throughout. Wang et al. (1991) suggested that a larger window-size was appropriate when there was a non-linear response and has shown that using a window wider than the actual wavelength shift can improve performance for PDS. We did not find this to be the case for our data.

11.6 Selection of standardisation samples

As we mentioned in chapter 3, two methods are used in the literature for selecting standardisation samples in an experimental situation where spectra for the



(a)



(b)

Figure 11.4: Graphs showing regression coefficients for PDS for NIR data, using a) Bayes with plug-in parameters b) DLM. Intercept-blue, central coefficient-red, side coefficients green, cyan.

entire calibration set are available on both the slave and master instruments. We used the method proposed by Kennard and Stone (1969) and used by Wang et al. (1992) and Wang et al. (1991) because it seemed likely to provide a more appropriate standardisation set for our methods than the standardisation set used by Bouveresse et al. (1994) and Bouveresse et al. (1995). It became clear from our results that there was a problem with our standardisation set of eight samples for the NIR data in that the additional three samples sometimes made results worse than using only five samples. Also, as mentioned previously, because the standardisation samples are in some sense the most extreme in the calibration set they are not the best for selecting parameter values. The alternative to selecting parameters that minimise RMSEP on the standardisation set is to use a separate set of spectra that are more typical of the prediction set. This means that more samples are needed for standardisation.

The selection of samples for standardisation is clearly an area that needs further investigation.

11.7 Conclusion

In this thesis we have considered two widely used standardisation methods, PDS and SW. Problems with both these methods appear to be due to lack of information at certain parts of the spectra and the correlations between absorbances at adjacent wavelengths. We have shown that Bayesian shrinkage and smoothing can, to some extent solve these problems. A criticism that could be levelled against our solutions to standardisation methods based on the PDS model is that because of the effect of correlation they fail to correct the perceived error and still tend to overfit. The SW method aims to correct the wavelength shift separately, but the wavelength shift correction step does not work well.

In an alternative approach we tried to improve the wavelength shift correction and then use Bayesian techniques based on a univariate model to derive

the regression coefficients for the absorbance correction. Smoothing the correlation coefficients used to estimate the wavelength shift function using a Gaussian filter was considerably better than the original SW method but still performed less well than methods based on the PDS model. Using wavelets to model the wavelength shift function worked much better and in fact, simple linear regression to correct the absorbance shift, after using the wavelet based wavelength correction was sufficient to produce good results. This last method seems to us to be the best of the methods we tested. It is fairly fast, is easy to use, parameter values are not critical and results are good. Using wavelet bases other than the Haar wavelet basis might improve performance. Other alteratives are to apply shrinkage and smoothing of the regression coefficients, estimating the parameters using ICM or simulated annealing or a Bayesian approach using MCMC, combining estimation of the wavelength shift in the wavelet domain with Bayesian shrinking and smoothing of the regression coefficents.

The main focus of this thesis has been on standardisation where there is a wavelength shift, but in chapter 10 we describe transfer by orthogonal projection, a method that corrects only for an absorbance shift. One conclusion that we can draw is that where there is known to be no wavelength shift, transfer by orthogonal projection is likely to be the best choice of standardisation method.

Appendix A

Proof that as $k \to \infty$ the scores for PCR are separated into scores for Q and those for $X - XQ^TQ$.

Let X be the $n \times w$ matrix of spectral responses on the master instrument. Let Q_0 be the $r_0 p \times w$ matrix of spectral responses of p samples on r_0 different instruments. From Q_0 we derive Q by deleting any dependent rows and pre-multiplying by an appropriate (lower triangular) matrix so that its rows are orthonormal.

The scores of $\begin{pmatrix} kQ \\ X \end{pmatrix}$ are the eigenvectors of $\begin{pmatrix} kQ \\ X \end{pmatrix} \begin{pmatrix} kQ^T & X^T \end{pmatrix}$ and hence satisfy

$$\begin{pmatrix} kQ \\ X \end{pmatrix} \begin{pmatrix} kQ^T & X^T \end{pmatrix} e = \lambda I e.$$

Let

$$M = \begin{pmatrix} I_r/k & O_{r,n} \\ -XQ^T/k & I_n \end{pmatrix}$$

where $O_{r,n}$ is an $r \times n$ matrix of zeros and I_r, I_n are identity matrices. Then

$$M \begin{pmatrix} kQ \\ X \end{pmatrix} \begin{pmatrix} kQ^T & X^T \end{pmatrix} M^T = \begin{pmatrix} kQ \\ R \end{pmatrix} \begin{pmatrix} kQ^T & R^T \end{pmatrix}, \qquad \text{(A.1)}$$

where $R = (X - XQ^TQ)$.

A.1 has eigenvectors of the form $\begin{pmatrix} \epsilon_1 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ \epsilon_2 \end{pmatrix}$

This follows since there exist matrices U_1 $(r \times r)$ and U_2 $(n \times n)$ such that $U_1 Q Q^T U_1^T$ and $U_2 R R^T U_2^T$ are diagonal matrices and if

$$U = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}$$

$$U \begin{pmatrix} kQ \\ R \end{pmatrix} \begin{pmatrix} kQ^T & R^T \end{pmatrix} U^T = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} kP \\ R \end{pmatrix} \begin{pmatrix} kQ^T & R^T \end{pmatrix} \begin{pmatrix} U_1^T & 0 \\ 0 & U_2^T \end{pmatrix}$$

$$= \begin{pmatrix} k^2\Lambda_1^2 & 0 \\ 0 & \Lambda_2^2 \end{pmatrix}$$

so the eigenvectors of $\begin{pmatrix} kQ \\ R \end{pmatrix} \begin{pmatrix} kQ^T & R^T \end{pmatrix}$ are the rows of U, with eigenvalues $k^2\lambda_i^2, i = 1...r$ and $\lambda_j^2, j = r+1...r+n$

i.e.

$$\begin{pmatrix} kQ \\ R \end{pmatrix} \begin{pmatrix} kQ^T & R^T \end{pmatrix} \begin{pmatrix} e_i \\ 0 \end{pmatrix} = k^2\lambda_i^2 \begin{pmatrix} e_i \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} kQ \\ R \end{pmatrix} \begin{pmatrix} kQ^T & R^T \end{pmatrix} \begin{pmatrix} 0 \\ e_j \end{pmatrix} = \lambda_j^2 \begin{pmatrix} 0 \\ e_j \end{pmatrix}$$

so from A.1

$$\begin{pmatrix} kQ \\ X \end{pmatrix} \begin{pmatrix} kQ^T & X^T \end{pmatrix} M^T \begin{pmatrix} e_i \\ 0 \end{pmatrix} = k^2\lambda_i^2 M^{-1} \begin{pmatrix} e_i \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} kQ \\ X \end{pmatrix} \begin{pmatrix} kQ^T & X^T \end{pmatrix} M^T \begin{pmatrix} 0 \\ e_j \end{pmatrix} = \lambda_j^2 M^{-1} \begin{pmatrix} 0 \\ e_j \end{pmatrix}$$

Now

$M = \begin{pmatrix} I_r & 0 \\ -XQ^T/k & I_n \end{pmatrix} \to I_{r+n}$ as $k \to \infty$ so for large k the eigenvectors of $\begin{pmatrix} kQ \\ X \end{pmatrix} \begin{pmatrix} kQ^T & X^T \end{pmatrix}$ and hence the scores of $\begin{pmatrix} kQ \\ X \end{pmatrix} \to \begin{pmatrix} e_i \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ e_j \end{pmatrix}$.

Appendix B

Proof of equivalence of repeatability file and elimination of variation between instruments for multiple linear regression in the single sample case with $k \to \infty$.

Repeatability File

Let Q_0 be the $r_0 \times w$ matrix of spectral responses of a single sample on r_0 different instruments. Each row in Q_0 is centred by subtracting from each the mean of all samples in Q_0. From Q_0 we derive the $r \times w$ matrix Q by deleting any dependent rows and pre-multiplying by an appropriate (lower triangular) matrix so that its rows are orthonormal. Let X be the $n \times w$ matrix of spectral responses of n different samples (including the one used in the repeatability file) on the master instrument, and let y be the vector of reference values for these samples.

Our model is

$$y_o = \begin{pmatrix} kQ \\ X \end{pmatrix} \beta_k + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \tag{B.1}$$

where

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N(0, \Sigma),$$

$y_o = (0, ..., 0, y^T)^T$, is a vector whose first r entries are zero and whose other entries are given by y and $\Sigma = \sigma^2 I_{r+n}$. Here, for the purposes of the regression,

we are treating kQ as a set of spectra with zero reference values.

We pre-multiply each side of equation B.1 by

$$M = \begin{pmatrix} I_r/k & O_{r,n} \\ -XQ^T/k & I_n \end{pmatrix}$$

where $O_{r,n}$ is an $r \times n$ matrix of zeros and I_r, I_n are identity matrices, giving

$$My_o = \begin{pmatrix} P \\ X - XQ^TQ \end{pmatrix} \beta_k + \begin{pmatrix} \epsilon_1/k \\ \epsilon_2 - XQ^T\epsilon_1/k \end{pmatrix}$$

so that

$$y_o = \begin{pmatrix} Q \\ R \end{pmatrix} \beta_k + \begin{pmatrix} \epsilon_1/k \\ \epsilon_2 - XQ^T\epsilon_1/k \end{pmatrix} \tag{B.2}$$

where R and Q span orthogonal spaces.

Now there exist orthogonal matrices U_1, $r \times r$ and T, $w \times w$ such that

$$U_1QT^T = (\Lambda \quad 0)$$

where Λ is a diagonal matrix

Multiplying B.2 by

$$\begin{pmatrix} U_1 & 0 \\ 0 & I_n \end{pmatrix}$$

gives

$$\begin{pmatrix} 0 \\ y \end{pmatrix} = \begin{pmatrix} \Lambda & 0 \\ RT^T \end{pmatrix} T\beta_k + \begin{pmatrix} U_1\epsilon_1/k \\ \epsilon_2 - XQ^T\epsilon_2/k \end{pmatrix}$$

explicitly,

$$\begin{aligned} 0 &= \begin{pmatrix} \Lambda & 0 \end{pmatrix} \gamma_k + U_1\epsilon_1/k \\ y &= R\beta_k + \epsilon_2 - XQ^T\epsilon_1/k \end{aligned}$$

where $\gamma_k = T\beta_k$. Now $\epsilon_1/k \sim N(0, \sigma^2 I_r/k^2)$ and

$$\epsilon_2 - XQ^T\epsilon_1/k \sim N(0, (I_n + QX^TXQ^T/k^2)\sigma^2)$$

From the first of these equations we deduce that $\hat{\gamma}_{ki} = 0, i = 1, ..., r$.

Now

$$\begin{aligned} Q\hat{\beta}_k &= U_1^T(U_1QT^T)T\hat{\beta}_k \\ &= Q_1^T(\Lambda_1\ 0)\hat{\gamma}_k \\ &= 0 \end{aligned}$$

Since $Q\hat{\beta}_k = 0$ and we may conclude that the members of the repeatability file are given zero prediction by the least squares regression estimator for the model given by equation (B.1). Also if we let $k \to \infty$, the model gives

$$y = R\beta_k + \epsilon_2$$

where $\epsilon_2 \sim N(0, \sigma^2 I_n)$

Transfer by orthogonal projection

Using the same notation we have, in the single sample case, P, the subspace of between instrument variation, is identical with Q, so

$$X = XQQ^T + R$$

We omit the first term on the right hand side, which represents the between-instrument variation and regress on R so our model is

$$y = R\tilde{\beta} + \epsilon_2$$

giving the same equation for $\tilde{\beta}$ as for β_k in the repeatability file, as $k \to \infty$.

Bibliography

Abramovich, F., T. Sapatinas, and B. W. Silverman (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B 60*(4), 725–749.

Amit, Y. (1994). A nonlinear variational problem for image matching. *SIAM Journal of Scientific Computing 15*, 297–224.

Andrew, A. and T. Fearn (2004). Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemometrics and Intelligent Laboratory Systems 72*, 51–56.

Aykroyd, R. G. and K. V. Mardia (1996). An MCMC approach to wavelet warping. In K. V. Mardia, C. A. Gill, and I. L. Dryden (Eds.), *Image Fusion and Shape Variability Techniques*, pp. 129–140. Leeds: Leeds University Press.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society 53*, 330–418.

Ben-Gera, I. and K. Norris (1968). Direct spectrophotometer determination of fat and moisture in meat products. *Journal of Food Science 33(1)*, 64–67.

Berkson, J. (1969). Estimation of a linear function of a calibration line; consideration of a recent proposal. *Technometrics 11*(4), 649–660.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice

systems (with discussion). *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Besag, J. E. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B 48*, 259–302.

Bouveresse, E., C. Hartmann, D. L. Massart, I. R. Last, and K. A. Prebble (1996). Standardisation of near-infrared spectrometric instruments. *Analytical Chemistry 68*, 982–990.

Bouveresse, E. and D. L. Massart (1996). Standardisation of near-infrared spectrometric instruments: a review. *Vibrational Spectroscopy 11*, 3–15.

Bouveresse, E., D. L. Massart, and P. Dardenne (1994). Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: effects of different standardisation samples. *Analytica Chimica Acta 297*, 405–416.

Bouveresse, E., D. L. Massart, and P. Dardenne (1995). Modified algorithm for standardisation of near-infrared spectrometric instruments. *Analytical Chemistry 67*(8), 1381–1389.

Brown, P., N. D. Le, and J. V. Zidek (1993). Inference for a covariance matrix. In R. Freeman and A. F. M. Smith (Eds.), *Aspects of Uncertainty: a Tribute to D. V. Lindley*. Chichester: Wiley.

Brown, P. J. (1993). *Measurement, Regression and Calibration.* Oxford: Clarendon.

Brown, P. J. and T. Makelainen (1992). Regression, sequenced measurements and coherent calibration. In J. M. Bernado, J. O. Berger, D. A. P., and A. F. M. Smith (Eds.), *Bayesian Statistics, 4*, pp. 97–108. Oxford: Clarendon Press.

Brown, P. J., M. Vannucci, and T. Fearn (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B 60*(3), 627–641.

Burns, D. A. and E. W. Ciurczak (Eds.) (1992). *Handbook of Near-infrared Analysis*. New York: Marcel Dekker, Inc.

Clyde, M., G. Parmigiani, and B. Vidakovik (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika 85*(2), 391–401.

Cooley, J. W. and J. W. Tukey (1965). An algorithm for the machine computation of complex Fourier series. *Mathematics of Computation 19*, 297–301.

Dardenne, P. and R. Biston (1991). Standardisation procedure and NIR network. In R. Biston and N. Bartiaux-Thill (Eds.), *Proceedings of the Third International Conference on Near Infrared Spectroscopy*, pp. 655–662. Gembloux, Belgium: Agricultural Research Centre Publishing.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics 41*, 909–996.

Dawid, A. P. (1981). Some matrix variate distribution theory: notational considerations and a Bayesian application. *Biometrica 68*, 265–274.

de Noord, O. E. (1994). Multivariate calibration standardisation. *Chemometrics and Intelligent Laboratory Systems 25*, 85–97.

Dean, T. and T. Isaksson (1993a). Standardisation. What it is and how it is done? Part 1. *NIR News 4*(2), 8–9.

Dean, T. and T. Isaksson (1993b). Standardisation. What it is and how it is done? Part 2. *NIR news 4*(4), 14–15.

Dean, T. and B. R. Kowalski (1996). Multivariate instrument standardisation: review of the state of the art. In S. D. Brown (Ed.), *Computer Assisted Analytical Spectroscopy*, pp. 175–187. Chichester: Wiley.

Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425–455.

Downie, T. R. (1997). *Wavelet Methods in Statistics.* Ph. D. thesis, University of Bristol.

Downie, T. R., L. Shepstone, and B. W. Silverman (1996). A wavelet based approach to deformable templates. In *Image Fusion and Shape Variability Techniques*, pp. 163–169. Leeds: Leeds University Press.

Duncan, D. B. and S. D. Horn (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association 67*, 815–821.

Fearn, T. (2000). On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems 50*, 47–52.

Fearn, T. (2001). Standardisation and calibration transfer for near infrared instruments: a review. *Journal of Near Infrared Spectroscopy 9*, 229–244.

Fearn, T. and A. M. C. Davies (2003). A comparison of Fourier and wavelet transforms in the processing of near infrared spectroscopic data: Part 1. Data compression. *Journal of Near Infrared Spectroscopy 11.*

Feudale, R. N., N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, and J. Ferre (2002). Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory Systems 64*, 181–192.

Fifield, F. W. and D. Kealey (1995). *Principles and Practice of Analytical Chemistry* (4 ed.). London: Blackie.

Geladi, P., D. McDougall, and H. Martens (1985). Linearisation and scatter correction for near infrared reflectance spectra of meat. *Applied Spectroscopy 39*, 491–500.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Speigelhalter (Eds.), *Markov chain Monte Carlo in Practice*. London: Chapman and Hall.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactionson Pattern Analysis and Machine Intelligence 6*, 721–741.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

Goldstein, M. (1976). Bayesian analysis of regression problems. *Biometrika 63*, 51–58.

Goldstein, M. and A. F. M. Smith (1974). Ridge-type estimators and regression analysis. *Journal of the Royal Statistical Society, Series B 36*, 284–291.

Grenander, U. (1970). A unified approach to pattern analysis. *Advances in Computers 10*, 175–216.

Hardy, C. L., G. R. Rippke, C. R. Hurburgh Jr, and T. J. Brumm (1996). Calibration and field standardisation of Foss Grainspec analysers for corn and soyabeans. In *Near Infrared Spectroscopy: The Future Waves*, pp. 132–141. Chichester: NIR Publicatons.

Harrison, P. J. and C. F. Stevens (1976). Bayesian forecasting. *Journal of the Royal Statistical Society, Series B 38*, 205–247.

Harvey, A. C. and G. D. A. Phillips (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika 66*, 49–58.

Hastings, W., K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistical Simulation and Computation 17(2)*, 581–607.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics 12*, 55–67.

Hurn, M. and C. Jennison (1995). A study of simulated annealing and a revised cascade algorithm for image reconstruction. *Statistics and Computing 5*, 175–190.

Jubb, M. and C. Jennison (1991). Aggregation and refinement in binary image restoration. In Possolo (Ed.), *Spatial Statistics and Imaging*, pp. 150–162. Institute of Mathematical Sciences, Lecture Notes, Hayward.

Kalman, R. E. (1963). New methods in Wiener filtering theory. In B. J. L. and F. Kozin (Eds.), *Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability*, pp. 270–388.

Kennard, R. W. and L. A. Stone (1969). Computer aided design of experiments. *Technometrics 11*(1), 137–148.

Kubelka, P. and F. Munk (1931). Ein Beitrag zur Optik der Farbanstriche. *Zeitschrift fur technische Physik 12*, 593–604.

Lambert, J. M. (1760). Photometria sive de mensura et grodibus luminis colorum et umbrae augustea vindelicorum.

Lindley, D. V. and A. F. M. Smith (1971). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B 34*(1), 1–41.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 11*, 674–693.

Mark, H. and J. Workman Jr (1988). A new approach to generating transferable calibrations for quantitative near-infrared spectroscopy. *Spectroscopy 3(11)*, 28–36.

Marquardt, D. (1970). Generalised inverses, ridge regression, biased linear estimation and non-linear estimation. *Technometrics 12*, 591–612.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1952). Equation of state calculations by fast computing machine. *Journal of Chemical Physics 21*, 1087–1091.

Naes, T., T. Isaksson, T. Fearn, and T. Davies (2002). *Multivariate Calibration and Classification.* Chichester: NIR Publications.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalised linear models. *Journal of the Royal Statistical Society, Series A 135*, 370–384.

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis.* Boston: Birkhäuser.

Osborne, B. G. and T. Fearn (1983). Collaborative evaluation of universal calibrations for the measurement of protein and moisture in flour by near infrared reflectance. *Journal of Food Technology 18*, 453–460.

Osborne, B. G., T. Fearn, and P. H. Hindle (1993). *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis* (Second Edition ed.). Harlow, Essex: Longman Scientific and Technical.

Park, K., Y. Ko, H. Lee, C. Jun, H. Chung, and M. Ku (2001). Near-infrared spectral data transfer using independent standardisation samples: a case study on the trans-alkylation process. *Chemometrics and Intelligent Laboratory Systems 55*, 53–65.

Roger, J.-M., F. Chauchard, and V. Bellon-Maurel (2003). EPO and PLS external parameter orthogonalisation of PLS application to temperature-

independent measurement of sugar content of intact fruits. *Chemometrics and Intelligent Laboratory Systems 66*, 191–204.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry 36*, 1627–1639.

Shenk, J. S. and M. O. Westerhaus (1989). U.s. patent. *4866644*.

Shenk, J. S. and M. O. Westerhaus (1991). New standardisation and calibration procedures for NIRS analytical systems. *Crop Science 31*, 1694–1696.

Stone, M. and R. J. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares squares and principal components regression. *Journal of the Royal Statistical Society, Series B 52(2)*, 237–269.

Tiao, G. C. and A. Zellner (1964). On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society, Series B 26*, 277–285.

Tillmann, P., T.-C. Reinhardt, and C. Paul (2000). Networking of near-infrared spectroscopy instruments for rapeseed analysis: a comparison of different procedures. *Journal of Near Infrared Spectroscopy 8*, 101–107.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Chichester: John Wiley and sons.

Walczak, B., E. Bouveresse, and D. L. Massart (1997). Standardisation of near-infrared spectra in the wavelet domain. *Chemometrics and Intelligent Laboratory Systems 36*, 41–51.

Wang, Y., T. Dean, and B. R. Kowalski (1995). Additive background correction in multivariate instrument standardisation. *Analytical Chemistry 67*, 2379–2385.

Wang, Y. and B. R. Kowalski (1992). Calibration transfer and measurement stability of near-infrared spectrometers. *Applied Spectroscopy 46*, 764–771.

Wang, Y., M. J. Lysaght, and B. R. Kowalski (1992). Improvement of multivariate calibration through instrument standardisation. *Analytical Chemistry 64*, 562–564.

Wang, Y., D. J. Veltkamp, and B. R. Kowalski (1991). Multivariate instrument standardisation. *Analytical Chemistry 63*, 2750–2756,.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models.* New York: Springer.

West, M., J. Harrison, and H. S. Migon (1985). Dynamic generalised linear models and Bayesian forecasting. *Journal of the American Statistical Association 80*, 73–83.

Westerhaus, M. O. (1991). Improving repeatability of NIR calibrations across instruments. In R. Biston and N. Bartiaux-Thill (Eds.), *Proceedings of the Third International Conference on Near Infrared Spectroscopy*, pp. 671–674. Gembloux, Belgium: Agricultural Research Centre Publishing.

Wise, B. M. and N. B. Gallagher (1998). *PLS Toolbox.* Eigenvector Research, Inc.

Wold, S., S. H. Antii, F. Lindgren, and J. Ohman (1998). Orthogonal signal correction of near infraed spectra. *Chemometrics and Intelligent Laboratory Systems 44*, 175–185.

Yoon, J., B. Lee, and C. Han (2002). Calibration transfer of near-infrared spectra based on compression of wavelet coefficients. *Chemometrics and Intelligent Laboratory Systems 64*, 1–14.